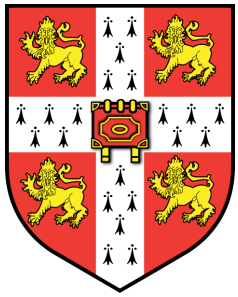
Essays in Econometrics



Ashish Patel
University of Cambridge
Christ's College

This dissertation is submitted for the degree of
Doctor of Philosophy

Faculty of Economics

July 2018

Essays in Econometrics

Ashish Patel

Summary

The thesis comprises four papers in theoretical econometrics: two papers on the subject of robust estimation and inference in moment condition models, one on semiparametric estimation in the presence of missing data, and one on survival analysis with competing risks data.

The first chapter considers estimation of moment condition models when some data are missing. The inverse probability tilting (IPT) estimator of Graham et al. [2] re-weights fully observed data to account appropriately for missingness. This paper considers a generalisation of the IPT estimator that allows for more flexible nuisance parameter estimation. It is shown that an IPT estimator with nonparametrically-estimated generated regressors retains some key asymptotic efficiency and robustness properties. A simulation study illustrates that these robustness properties allow IPT estimators to be insensitive to the choice of tuning parameter.

The second chapter concerns semiparametric moment condition models where the parameter of interest is described by one set of moment restrictions, while nuisance functions are identified from another set of moment restrictions. A two-step GEL-weighted estimator, a generalisation of Hellerstein and Imbens [3] and Bravo [1] to the semiparametric setting with estimated nuisance functions, is proposed that guarantees an efficiency gain arising from exploiting auxiliary moment restrictions that may involve nonparametric components. It is shown that in order to achieve this, moment restrictions generally need to be adjusted to account for first-stage nuisance estimation of nonparametric components. The theory is applied to a semiparametric missing data model where it is shown that the two-step GEL-weighted estimator possesses good efficiency and robustness properties when nuisance models are misspecified.

The third chapter represents my contribution on a project led by James Wason and Chien-Ju Lin of the MRC Biostatistics Unit. The paper focuses on time-to-event studies with several possible causes of event for each individual. When these competing risks are mutually dependent and only information on the time-to-first-event is available, marginal survival functions for each risk cannot be identified. Copula-Graphic estimators (Zheng and Klein [5]) exploit information on the dependence structure between risks to return consistent estimators. The paper derives asymptotic results for a class of parametric Copula-Graphic estimators, allowing for the construction of asymptotic confidence intervals for marginal survival functions. The performance of these confidence intervals is investigated in a simulation study.

The final chapter (co-authored with my supervisor, Richard Smith) considers methods to investigate the validity of over-identified moment restrictions when violations may occur only in small subgroups of the population. Hansen's J-test and likelihood-based variants aim to have non-trivial power against a wide range of alternatives, whereas power against particular forms of heterogeneity or parameter instability are often of concern. The paper addresses this issue by providing concentration inequalities designed to detect patterns of model misspecification. The associated bounds can be used to identify subsets of individual characteristics that are not consistent with the moment restrictions. These results are applied to show the consistency of goodness-of-fit statistics (Ramalho and Smith [4]) with data-dependent partitions.

References

[1] Francesco Bravo. Efficient M-estimators with auxiliary information. *Journal of Statistical Planning and Inference*, 140(11):3326–3342, 2010.

[2] Bryan S. Graham, Cristine Campos De Xavier Pinto, and Daniel Egel. Inverse probability tilting for moment condition models with missing data. *Review of Economic Studies*, 79(3):1053–1079, 2012.

[3] Judith K. Hellerstein and Guido W. Imbens. Imposing Moment Restrictions from Auxiliary Data by Weighting. *The Review of Economics and Statistics*, 81(1):1–14, 1999.

[4] Joaquim J S Ramalho and Richard J. Smith. Goodness of Fit Tests for Moment Condition Models. 2006.

[5] Ming Zheng and John P. Klein. Estimates of marginal survival for dependent competing risks based on an assumed copula. *Biometrika*, 82(1):127–138, 1995.

Declaration

This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Summary.

This dissertation is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution.

This dissertation does not exceed the 60,000 words, in accordance with the requirements set forth by the Economics Degree Committee.

Ashish Patel
Cambridge, July 2018

Acknowledgments

It has been an honour to work on this research under the supervision of Richard Smith. I thank him for his guidance and support which have been invaluable.

I am grateful to the Faculty of Economics and Christ's College for providing a wonderful research environment and for their support on many matters. It has been a great pleasure to learn from all econometricians at the department; I thank them for their detailed feedback during presentations, and for the many discussions and comments that have shaped the dissertation.

Part of the research for this dissertation has been undertaken while based at the MRC Biostatistics Unit, University of Cambridge. I thank James Wason and Paul Newcombe for providing that opportunity.

I thank Alexey Onatskiy (internal examiner) and Frank Windmeijer (external examiner) for their detailed comments and suggestions which have greatly improved the dissertation.

I am grateful for the generous financial support received from an ESRC Studentship. I also thank Christ's College and the Keynes Fund for financial assistance.

In loving memory of my father, Pravin Patel.

Contents

1. Inverse Probability Tilting with Generated Regressors for Missing Data Models 7

2. Robust Estimation Using Auxiliary Semiparametric Information 52

3. Copula-Graphic Inference with Cause-specific Hazard Models 97

4. Large Deviation Bounds for Inference in Moment Condition Models 133

Ph.D. Thesis: Chapter One

Inverse Probability Tilting with Generated Regressors for Missing Data Models

Ashish Patel*
University of Cambridge

Abstract

This paper considers the estimation of moment condition models when some data are missing. In this setting, existing inverse probability weighting estimators compete on the basis of efficiency and robustness. The inverse probability tilting (IPT) estimator of Graham et al. [12] has many attractive properties relative to those estimators. This paper studies an extension of the IPT estimator that allows for more flexible nuisance parameter estimation. It is shown that an IPT estimator with generated regressors achieves the semiparametric efficiency lower bound. In the presence of auxiliary moment conditions that are correlated with the IPT estimating equations, generalised empirical likelihood-weighting provides an efficiency gain. The IPT estimator considered here retains a double robustness property that is not shared by alternative one-step approaches that are based on stacking all moment conditions. A simulation study shows the IPT estimator with generated regressors may offer an improvement in finite sample properties over alternative estimators suggested in the literature, and illustrates how its double robustness property enables IPT estimators to be robust to the choice of tuning parameters.

Keywords: Moment Condition Models, Double Robustness, Missing Data, Propensity Score Estimation

*I am grateful for detailed feedback and comments from Oliver Linton, Richard Smith and Melvyn Weeks. I thank Shaun Seaman and Haihan Tang for helpful discussions. Financial support from an ESRC Studentship Award is gratefully acknowledged.

1 Introduction

Researchers are often faced with a problem of missing data. For example, survey participants may provide limited information when completing questionnaires. Policy evaluation or analysis of clinical trials must often overcome the issue of participant dropout, in which those enrolled in a certain treatment programme may be required to report weekly on response variables. Moreover, the literature of treatment effects estimation is closely related with that on missing data since the unobservable counterfactual outcomes can also be regarded as missing data.

In some circumstances, the standard estimation and inference methods remain valid even if some data are missing. However, in general, these methods may need to be adapted to account for missing data. This paper considers estimation when data are *missing at random* (MAR); see Rubin [35]. Under the MAR assumption, the event that a response is missing is independent of its value given the fully-observed covariates. Models based on this assumption have proved very popular for handling missing data. A widely-used method which yields consistent parameter estimates in this missing data set-up is *inverse probability weighting* (IPW), first proposed by Horvitz and Thompson [21]. IPW estimators re-weight fully observed data to account appropriately for missingness.

The subject of this paper is estimation in unconditional moment restriction models with missing data. For such models, many existing IPW estimators compete on the basis of efficiency and robustness. The semiparametric efficiency lower bound for the missing data model considered here is characterised in Robins et al. [34]. Estimators based on the efficient influence function are contained in the class of *augmented IPW* estimators. For augmented IPW estimators, the estimation of two nuisance functions is required. However, only one of the models specified for the nuisance functions is required to be correct for augmented IPW estimators to be consistent. This is the *double robustness* property that has become very popular in applied work; see, for example, Scharfstein et al. [38]. As Bang and Robins [3] note, the analyst is given "two chances, instead of only one, to make a valid inference".

Graham et al. [12], henceforth, GPE, recently developed the inverse probability tilting (IPT) estimator. IPT estimators have many attractive features as compared with augmented IPW estimators, in terms of both asymptotic and finite sample properties. Given correct specification of the nuisance functions, IPT estimators attain the semiparametric efficiency lower bound of Robins et al. [34], and have better higher-order bias properties than the class of augmented IPW estimators. IPT estimators are also doubly robust, allowing consistent estimation of the parameter of interest under some forms misspecification of the nuisance functions. However, when compared to augmented IPW estimators, the estimator studied in GPE imposes relatively restrictive assumptions on specifications of the *propensity score* (conditional probability that data are missing given the covariates) and the *conditional expectation function* (the conditional expectation of moment functions given the covariates). This paper analyses the properties of IPT estimators when such assumptions are weakened.

Of course, the MAR assumption may only be plausible if the propensity score model involves conditioning on a high number of observable characteristics. Consequently, approaches which allow more flexible propensity score specifications are of interest. For example, Belloni et al. [4] consider properties of various treatment effect estimators when machine learning is used for first-step propensity score estimation. Also, Hirano et al. [20] show that an IPW estimator weighted by a logit propensity score with approximating functions that span the space of observables is semiparametrically efficient.

To make an IPT approach more flexible and comparable with existing estimators, this paper considers IPT estimators that allow for generated regressors in the first step. Two cases are considered: one in which the generated regressors are nonparametrically-determined, and the other in which the approximating functions may be parameterised by a finite-dimensional unknown vector. The latter approach induces a further question of how best to utilise the auxiliary information that may describe the finite-dimensional parameter. A potential trade-off between efficiency and robustness is considered.

The results of this paper show that an IPT estimator with generated regressors (IPTGR) achieves the same asymptotic variance as the efficiency lower bound of Robins et al. [34]; consequently, there is no loss of efficiency relative to competing estimators even with plug-in estimators entering the propensity score. It is also shown that IPTGR preserves the double robustness property, requiring only one of the two nuisance models to be specified correctly for consistency. In presence of auxiliary information, for example, containing population information, as in Hellerstein and Imbens [18], an estimator based on re-weighting IPT estimating equations by generalised empirical likelihood weights entails an efficiency gain while also retaining a double robustness property that guards against some forms of misspecification of nuisance functions. Finally, an alternative one-step estimation approach based on all moment conditions jointly that describe the missing data model is shown not to share this double robustness property.

The next section introduces the moment condition model and missing data set-up, and compares existing IPW estimators. Section 3 discusses the IPTGR estimator and presents its properties under correct specification and under different forms of misspecification. Section 4 presents a simulation study illustrating the use of IPTGR and makes comparisons with existing IPW and imputation estimators. Section 5 concludes. All proofs are given in the Appendix.

The following abbreviations are used. $\xrightarrow{p}$: converges in probability to; $\xrightarrow{d}$: converges in distribution to; T : the triangle inequality; M: the Markov inequality; CS: the Cauchy-Schwarz inequality; UWL: the uniform weak law of large numbers (for example, Lemma 2.4 of Newey and McFadden [27]); WLLN: the weak law of large numbers; CLT: the central limit theorem; LIE: the law of iterated expectations; LHS: left hand side; RHS: right hand side; w.p.1: with probability one; w.p.a.1: with probability approaching one; $||.||$ is the Euclidean norm.

2 Moment Condition Model with Missing Data

The moment conditions set-up and notation considered here closely follows GPE. Let $z = (y, x) \in \mathcal{Z}$ be a vector of variables where y is a d_y-dimensional vector; possibly missing, and x a fully observed d_x-dimensional vector of covariates. Of central concern is the estimation of γ_0, the d_γ-dimensional unknown parameter which uniquely satisfies the moment condition

$$\mathbb{E}[\psi(z, \gamma_0)] = 0, \tag{2.1}$$

where $\psi(z, \gamma)$ is a known d_ψ-dimensional vector of the data z and the unknown parameter γ. The identification of γ_0 requires that $d_\psi \geq d_\gamma$. However, for simplicity it is assumed that γ_0 is just-identified, that is, $d_\psi = d_\gamma$. The extension to the over-identified case, $d_\psi > d_\gamma$, is straight-forward.

Consider a situation in which y is not always observed. Let d be a binary indicator variable for whether or not y is observed, i.e., $d = 1$ if y is observed, $d = 0$ otherwise. For consistent estimation of γ_0, it is the relationship between d and y that determines whether missing data on y is ignorable, insurmountable or something in between. If $\{z_i, d_i\}_{i=1}^n$ is a sample of n units, the researcher only observes $\{x_i, d_i, dy_i\}_{i=1}^n$ where for each unit $(i = 1, ..., n)$, the vector dy_i equals y_i if $d_i = 1$, and has a missing entry if $d_i = 0$.

2.1 Missing at Random and Inverse Probability Weighting

2.1.1 Missing at Random (MAR)

The implications missing data have for the validity of the usual estimation and inference procedures depend on the restrictions imposed by the missing data. Many types of missing data schemes have been studied. Rubin [36] and Little and Rubin [24] discuss a classification system for missing data assumptions that has come to define convention.

There are cases when the occurence of missing data can be ignored with the usual estimation methods valid based on the completely observed units $\{dy_i, dx_i\}_{i=1}^n$, where dx_i is defined analogously to dy_i above. One example is when data are missing completely at random (MCAR), i.e., y is MCAR if the probability of observing y does not depend on values of y or x. For example, data is MCAR if the conditional probability that y is missing is constant, $\mathbb{P}(d = 1|y, x) = c$, for some $c > 0$. Wooldridge [43] discusses conditions under which consistent estimation need not account for missing data.

If the probability of observing y depends on the values of y, then y is said to be missing not at random (MNAR). If y is MNAR, there is an identifiability issue that stems from the probability $\mathbb{P}(d = 1|y)$ not being estimatable from observable data because if $d = 0$, then y is unobserved (Robins [33]). Without further restrictions, γ_0 from (2.1) cannot be consistently

estimated; for further discussion see Tsiatis ([40], p. 140-3). An example of such a restriction is the Heckman correction (Heckman [17]) which solves the MNAR problem by modelling the probability that data are missing (or from the selection bias perspective, the probability that an individual finds work) in the first stage.

This paper considers the case when y is missing at random (MAR). That is, (i) the reason y is missing depends on some fully observed variables x; (ii) conditional on x, the value of y has no additional effect on the probability that y is missing,

$$d \perp y|x, \tag{2.2}$$

i.e., conditional on x, y is independent of d. Although MAR is slightly more general than selection on observables, in most practical applications it is the same.

The MAR assumption is likely to be satisfied with experimental level data. For example, if a treatment rule is decided before treatment, then the probability that a person is treated is independent of treatment outcomes, and only dependent on the fully observed covariate information x determining the treatment rule.

The MAR assumption has also been popular in observational level studies. For example, there may exist fully observed surrogate variables closely related to missing outcomes as in Chen et al. [7]. To investigate voting behaviour in US general elections, they consider eligible individuals who did not vote as constituting a missing data problem. Post-election data in which non-voters were asked their preferred candidates are termed *surrogate outcomes*. A MAR restriction is then imposed for estimation, in particular, the assumption that the event an individual did not vote was independent of who they would have voted for, given their surrogate outcome.

In some cases, to render the MAR assumption more plausible, x may be a high-dimensional vector, see, for example, Belloni et al. [4]. The MAR assumption would then imply that after controlling for various combinations of a high-dimensional vector of covariates x, y is independent of d.

2.1.2 Inverse Probability Weighting (IPW)

Popular methods that deal with the problem of MAR data include likelihood and imputation methods, see, for example, Rubin [37] and Wang and Rao [42], where the distribution of y conditional on x is estimated. Arguably, the most popular method is IPW which reweights all fully observed units by the inverse of the probability of selection.

Let $p_0(x) = \mathbb{P}(d = 1|x)$ be the propensity score, i.e., the conditional probability that y are missing given x. An IPW estimator for γ_0 based on the moment condition model (2.1) is given

by the solution $\hat{\gamma}_{IPW}$ to

$$\frac{1}{n}\sum_{i=1}^{n}\frac{d_i\psi(z_i,\hat{\gamma}_{IPW})}{p_0(x_i)}=0.$$

This approach results in a consistent estimator for γ_0 through the following argument. Suppose $\hat{\gamma}_{IPW}$ converges to some value γ_1 as $n\to\infty$. Then, under regularity conditions see, for example, Newey and McFadden [27], Lemma 2.4, by a uniform law of large numbers, $n^{-1}\sum_{i=1}^{n}(d_i\psi(z_i,\hat{\gamma}_{IPW})/p_0(x_i))\xrightarrow{p}\mathbb{E}[d\psi(z,\gamma_1)/p_0(x)]$. Now,

$$\begin{aligned}\mathbb{E}\Big[\frac{d\psi(z,\gamma_1)}{p_0(x)}\Big] &= \mathbb{E}\Big[\mathbb{E}\Big[\frac{d\psi(z,\gamma_1)}{p_0(x)}\Big|y,x\Big]\Big]\\ &= \mathbb{E}\Big[\frac{\mathbb{E}[d|y,x]}{p_0(x)}\psi(z,\gamma_1)\Big]\\ &= \mathbb{E}\Big[\frac{\mathbb{E}[d|x]}{p_0(x)}\psi(z,\gamma_1)\Big]\\ &= \mathbb{E}[\psi(z,\gamma_1)],\end{aligned}$$

where the first equality follows from the law of iterated expectations, the second since $z=(y,x)$, the third from the MAR assumption (2.2), and the fourth from $\mathbb{E}[d|x]=\mathbb{P}(d=1|x)=p_0(x)$. Under standard identification conditions guaranteeing the uniqueness of γ_0, γ_1 must equal γ_0, that is, the IPW estimator is consistent.

The general approach of re-weighting observed data by the inverse probability of selection was first proposed by Horvitz and Thompson [21], and several developments of IPW methods have been used in applications across many disciplines including health sciences (Hernán et al. [19]) and economics (for example, accounting for non-responses in surveys, Chen et al. [8]).

2.2 Model assumptions

In this subsection, the assumptions maintained for asymptotic inference are discussed. They are similar to Assumptions 2.1-2.5 and 3.1 of GPE.

Assumption 2.1 (Identification). (i) $\mathbb{E}[\psi(z,\gamma)]=0$ *uniquely at* $\gamma=\gamma_0\in\Gamma\subset\mathbb{R}^{d_\gamma}$**; (ii)** Γ *is compact and* $\gamma_0\in int(\Gamma)$**; (iii)** $\psi(z,\gamma)$ *is continuous at each* $\gamma\in\Gamma$ *with probability one, and continuously differentiable in a neighbourhood* Γ_0 *of* γ_0**; (iv)** $\sup_{\gamma\in\Gamma}||\psi(z,\gamma)||\leq b_\psi(z)$ *for all* $z\in\mathcal{Z}$ *where* $b_\psi(z)\geq 0$ *is such that* $\mathbb{E}[b_\psi(z)]<\infty$**; (v)** $\mathbb{E}[||\psi(z,\gamma_0)||^2]$ *is finite,* $\Psi=\mathbb{E}[\nabla_\gamma\psi(z,\gamma_0)]$ *has full rank* d_ψ*, and* $\sup_{\gamma\in\Gamma}||\nabla_\gamma\psi(z,\gamma)||<b_\Psi(z)$ *for all* $z\in\mathcal{Z}$ *where* $b_\Psi(z)\geq 0$ *is such that* $\mathbb{E}[b_\Psi(z)]<\infty$.

Assumption 2.1 provides standard conditions for consistency and asymptotic normality for moment of moments estimation of γ_0 when an i.i.d. sample on a fully observed vector z is available. The conditions are collected from the hypotheses of Theorems 2.6 and 3.4 of Newey

and McFadden [27]. By Chamberlain [6], the semiparametric asymptotic variance lower bound of any estimator of γ_0 is given by $(\Psi'(\mathbb{E}[\psi(z,\gamma_0)\psi(z,\gamma_0)'])^{-1}\Psi)^{-1}$.

Assumption 2.2 (Random sampling). *For $z = (y, x)$, $\{z_i, d_i\}_{i=1}^n$ is an i.i.d sequence, from which only $\{d_i, x_i, dy_i\}_{i=1}^n$ is available to the researcher, where $dy = y$ if $d = 1$ and has a missing entry if $d = 0$.*

Assumption 2.2 formalises the above discussion concerning missingness of z. In particular, given a sample size of n, y_i is generally not observed for every $i = 1, ..., n$, although both d_i and x_i are fully observed for every $i = 1, ..., n$. When $d_i = 1$ data (d_i, y_i, x_i) are available, whereas when $d_i = 0$ only (d_i, x_i) are available.

Assumption 2.3 (Missing at random). *$d \perp y|x$, in particular, $\mathbb{P}(d = 1|y, x) = \mathbb{P}(d = 1|x)$.*

See Section 2.1.1 above for a discussion of the MAR assumption.

Assumption 2.4 (Overlap). *For $p_0(x) = \mathbb{P}(d = 1|x)$, $0 < \kappa \leq p_0(x) \leq 1$ for some $0 < \kappa < 1$ and for all $x \in \mathcal{X} \subset \mathbb{R}^{d_X}$.*

Assumption 2.4 is standard for propensity score matching estimation, allowing counterfactual analysis since, with probability one, there is no value of x for which y cannot be observed.

Assumption 2.5 (Propensity score model). *There is a unique $\delta_0 \in int(\mathcal{D})$ where $\mathcal{D} \subset \mathbb{R}^{d_r}$ is compact, and $r(x)$ a d_r-dimensional known vector of linearly independent functions of x such that $p_0(x) = G(r(x)'\delta_0)$ for all $x \in \mathcal{X}$, where $G : \mathbb{R} \to [0, 1]$ is a known function such that* **(i)** *$G(.)$ is strictly increasing and continuously differentiable,* **(ii)** *$\lim_{v\to-\infty} G(v) = 0$ and $\lim_{v\to\infty} G(v) = 1$, and* **(iii)** *$0 < \kappa \leq G(r(x)'\delta) \leq 1$ for all $\delta \in \mathcal{D}$ and $x \in \mathcal{X}$.*

Assumptions 2.1-2.4, and some regularity conditions for nonparametric estimation of the propensity score, are sufficient for consistency and asymptotic normality of $\hat{\gamma}_{IPW}$. Hahn [13] shows that estimators of γ_0 based on nonparametric estimation of the propensity score can be semiparametrically efficient. Consequently, Assumption 2.5 does not increase the precision with which γ_0 can be estimated. However, even if the propensity score is known, estimating the propensity score for IPW methods can lead to an efficiency gain. This so-called *propensity score puzzle* is explained by the observation that propensity score estimation incorporates information that is otherwise ignored by the ordinary IPW estimator as described in Section 2.1.2 above (also see Prokhorov and Schmidt [30] and Graham [11]).

Parametric models for the propensity score are often used due to small sample issues, and widely-used specifications such as the logit and probit models satisfy Assumption 2.5. Furthermore, the double robustness property allows consistent estimation of γ_0 even when the propensity score is misspecified. Such properties may make researchers more sanguine about imposing parametric restrictions for nuisance functions. The linear-in-parameters restriction of Assumption 2.5 is necessary for the construction of IPT estimators that exploit the corre-

lation between independent functions that enter the propensity score and the functions that describe the conditional expectation function; see Section 2.3.3 below.

2.3 IPW estimators

The intuition behind the consistency of the IPW method has been discussed in Section 2.1.2. Here, some developments of IPW estimation in general, and some other extensions closely related to the focus of this paper, are outlined.

2.3.1 Doubly robust IPW - Robins et al. [34]

The wide class of augmented IPW estimators are based on estimating equations resulting from the efficient influence function derived in Robins et al. [34]. A semiparametrically efficient estimator, based on Assumptions 2.1-2.4, of γ_0 is the solution $\hat{\gamma}_{RRZ}$ to the equation

$$\frac{1}{n}\sum_{i=1}^{n}\frac{d_i\psi(z_i,\hat{\gamma}_{RRZ})}{p_0(x_i)} - \Big(\frac{d_i}{p_0(x_i)} - 1\Big)q_0(x_i;\hat{\gamma}_{RRZ}) = 0, \tag{2.3}$$

where $q_0(x;\gamma_0) = \mathbb{E}[\psi(z,\gamma_0)|x]$ is the conditional expectation function. In general, both $p_0(x)$ and $q_0(x;\gamma_0)$ must be estimated and plugged into (2.3). While $p_0(x)$ is conceptually easy to estimate nonparametrically, if $\psi(z,\gamma_0)$ is highly nonlinear, nonparametric estimation of $q_0(x;\gamma_0)$ may be more challenging. Many IPW estimators therefore specify a working model for $q_0(x;\gamma_0)$; for example, Wang et al. [41] considers a partially linear model for $q_0(x;\gamma_0)$ when γ_0 represents a population mean. However, given a model specified for $q_0(x;\gamma_0)$ in addition to Assumptions 2.1-2.4, the estimator $\hat{\gamma}_{RRZ}$ may no longer be semiparametrically efficient; the calculation of the semiparametric asymptotic variance lower bound changes since there is now more information. Graham [11] provides efficiency bound calculations when $q_0(x;\gamma_0)$ can be modelled by a semiparametric conditional moment restriction.

Part of the attraction for specifying models for both $p_0(x)$ and $q_0(x;\gamma_0)$ is that estimation of γ_0 based on plug-in estimates in (2.3) is consistent as long as the model for least one of $p_0(x)$ and $q_0(x;\gamma_0)$ is correctly specified. This is the double robustness property.

2.3.2 Series logit IPW - Hirano et al. [20]

Constructing estimating equations based on the efficient influence function, as in (2.3), is one way of obtaining efficient estimators of γ_0. The series logit IPW estimator of Hirano et al. [20] is also efficient despite avoiding nuisance estimation of $q_0(x;\gamma_0)$. If $G(.)$ is the logit function, the functions $r(x)$ in Assumption 2.5 correspond to series approximating functions as in, for example, Newey [26]. As such, the Hirano et al. [20] estimator satisfies Assumption 2.5, $p_0(x) = G(r(x)'\delta_0)$, when the dimension of the linearly independent functions $r(x)$ is allowed

to increase with the sample size. The logit-series IPW estimator of γ_0 is the solution $\hat{\gamma}_{HIR}$ to the equation

$$\frac{1}{n}\sum_{i=1}^{n}\frac{d_i\psi(z_i,\hat{\gamma}_{HIR})}{G(r^s(x_i)'\hat{\delta}_s)}=0,$$

where $\hat{\delta}_s$ is an s sequence of estimated parameters and $r^s(x)=(r^1(x),...,r^s(x))$ an s sequence of approximating functions of x with $s\to\infty$ at an appropriate rate, see Hirano et al. ([20], p.1170).

Graham [11] shows that, under Assumptions 2.1-2.4, an efficient estimator exhausts the information provided by the moment conditions

$$\mathbb{E}\Big[\frac{d}{p_0(x)}\psi(z,\gamma_0)\Big] = 0 \tag{2.4}$$

$$\mathbb{E}\Big[\Big(\frac{d}{p_0(x)}-1\Big)\Big|x\Big] = 0, \tag{2.5}$$

with (2.5) nonparametrically identifying $p_0(x)$ and confirming the Hahn [13] result that efficient estimation is possible through a nonparametrically estimated propensity score. The logit-series estimator of $p_0(x)$ allows the estimated propensity score to contain independent functions of x that span the space of covariates $\mathcal{X}$. This approach allows the correlation between the moment function $\psi(z,\gamma)$ and functions of x that may be related to $q_0(x;\gamma_0)=\mathbb{E}[\psi(z,\gamma_0)|x]$ to be exploited for an efficiency gain, providing the intuition for why the Hirano et al. [20] estimator is semiparametrically efficient despite not being directly based on the efficient influence function.

2.3.3 Inverse probability tilting - Graham et al. [12]

In the study of treatment effects under experimental design, the rule that determines treatment, and hence the propensity score, is known. However, noting equation (2.5) and the results of Hirano et al. [20], it may be advantageous to 'over-specify' purposely the propensity score model such that other approximating functions are also used despite not being strictly necessary. This observation motivates the IPT estimator.

Suppose

$$q_0(x;\gamma_0):=\mathbb{E}[\psi(z,\gamma_0)|x]=\Pi_0^\star t^\star(x), \tag{2.6}$$

where $\Pi_0^\star$ is some unknown $d_\psi\times d_{t^\star}$ matrix, and $t^\star(x)$ is a known $d_{t^\star}$-vector of linearly independent functions of x which includes a constant as its first element. From Assumption 2.5, the d_r-vector of linearly independent functions of x that enter the propensity score is denoted by $r(x)$. Let $t(x)$ be the union of linearly independent functions of $r(x)$ and $t^\star(x)$, and write $t(x)=(r(x)',r^\star(x)')'$ where $r^\star(X)$ is the relative complement of $r(X)$ in $t^\star(X)$. Suppose the dimension of vector $t(X)$ is $d_t\le d_r+d_{t^\star}$.

The IPT estimator $\hat{\gamma}_{IPT}$ of γ_0 is the solution to the estimating equation (2.7)

$$\frac{1}{n}\sum_{i=1}^{n}\frac{d_i\psi(z_i,\hat{\gamma}_{IPT})}{G(t(x_i)'\hat{\delta})}=0, \tag{2.7}$$

given $\hat{\delta}$ as the solution to (2.8)

$$\frac{1}{n}\sum_{i=1}^{n}\Big(\frac{d_i}{G(t(x_i)'\hat{\delta})}-1\Big)t(x_i)=0. \tag{2.8}$$

Therefore, the IPT estimator is the solution to a simple method of moments problem. In the first step, (2.8), the propensity score model is potentially overfitted to include additional approximating functions that, although are not necessary to model $p_0(x)$, are correlated with $q_0(x;\gamma_0)$ potentially yielding more efficient estimation of γ_0 in the same spirit as Hirano et al. [20]. The propensity score parameter estimated in the overfitted model is the d_t-dimensional parameter $\delta_1=(\delta_0',0')'$ where δ_0 is the true propensity score parameter of Assumption 2.5, and 0 is a (d_t-d_r)-dimensional vector of zeros. Since at the true values $t(x)'\delta_1=r(x)'\delta_0+r^\star(x)'0$, $G(t(x)'\hat{\delta})$ is a consistent estimator of the propensity score if $\hat{\delta}\xrightarrow{p}\delta_1$.

Given an estimate of $p_0(x)=G(r(x)'\delta_0)$ from (2.8), in the second step, an estimator of γ_0 solves the method of moments problem (2.7). While such a two-step procedure might ordinarily incur an efficiency loss relative to joint estimation of (2.7) and (2.8), since (2.8) is just-identified by design, there is no loss of efficiency from sequential estimation, see, for example, Prokhorov and Schmidt [30] and Ackerberg et al. [1]. Furthermore, GPE shows that this two-step procedure is not only locally efficient, see Section 3.2.1, but also higher-order efficient relative to the class of augmented IPW estimators discussed in Section 2.3.1. Moreover it is doubly robust, that is, as long as at least one of (i) the conditional expectation function $\mathbb{E}[\psi(z,\gamma_0)|x]=\Pi_0^\star t^\star(x)$, or (ii) the propensity score model $p_0(x)=G(r(x)'\delta_0)$ is correctly specified, the IPT estimator of γ_0 based on (2.7) and (2.8) remains consistent.

Consider the following simple example. If the functions $r(x)$ entering the propensity score model are richer than the functions $t^\star(x)$ entering the conditional moment function there is no need to overfit the propensity score model. For example, if the propensity score involves quadratic elements $r(x)=(1,x,x^2)'$ but the conditional expectation function involves only linear functions $t^\star(x)=(1,x)'$, $\hat{\delta}$ solves (2.8), that is $\sum_{i=1}^n(d_iG(r(x_i)'\hat{\delta})^{-1}-1)r(x_i)/n=0$, and, under Assumptions 2.1-2.5, $\hat{\delta}\xrightarrow{p}\delta_0$.

If, on the other hand, the conditional moment function (2.6) is more complex, for example, $t^\star(x)=(1,x,x^2)'$, while the propensity score model satisfies (2.5) with $r(x)=(1,x)'$, then the IPT method overfits the propensity score model. Let $t(x)$ collect the linearly independent elements in $t^\star(x)$ and $r(x)$. In this case, $t(x)=(1,x,x^2)'$. Then $\hat{\delta}$ solves (2.8), that is, $\sum_{i=1}^n(d_iG(t(x_i)'\hat{\delta})^{-1}-1)t(x_i)/n=0$ and, under Assumptions 2.1-2.5, $\hat{\delta}\xrightarrow{p}\delta_1$ where $\delta_1=(\delta_0',0)'$, with δ_0 the true propensity score parameter of Assumption 2.5 with $r(x)=(1,x)'$.

In other words, an extra nuisance parameter has been estimated which has true value zero. The dimension of extra nuisance parameters increases one-for-one with independent elements in $t(x)$ that are not contained in the original propensity score elements $r(x)$.

2.3.4 IPW with auxiliary information - Chen et al. [7]

Chen et al. [7] is closely related to the focus of this paper and considers properties of IPW estimators when a model for the conditional expectation function $\mathbb{E}[\psi(z,\gamma_0)|x]$ is misspecified. Let $q(x,\eta_0)$ denote an approximation to $\mathbb{E}[\psi(z,\gamma_0)|x]$ known up to the finite-dimensional unknown parameter η_0. Consider the biased-corrected moment indicator vectors

$$\begin{aligned} g_1(d,x,\eta,\mu) &= \frac{d}{p(x)}(q(x,\eta)-\mu) \\ g_2(d,x,\eta,\mu) &= \frac{1-d}{1-p(x)}(q(x,\eta)-\mu), \end{aligned}$$

where $p(x)$ is the propensity score, and the unknown parameter μ_0 satisfies $\mu_0 = \mathbb{E}[q(x,\eta_0)]$. By LIE, each moment indicator has mean zero by construction whether or not $p(x)$ is correctly specified (cf. Chen et al. [7], p. 808).

Given a $\sqrt{n}$-consistent estimator for η_0, and a consistent estimator for the propensity score $p_0(x)$, an empirical likelihood procedure estimates the parameter μ from the moment conditions $\mathbb{E}[g_1(d,x,\eta_0,\mu_0)] = 0$ and $\mathbb{E}[g_2(d,x,\eta_0,\mu_0)] = 0$. The estimator $\hat{\gamma}_{CLQ}$ of γ_0 solves an IPW estimating equation that is re-weighted by empirical likelihood weights $\{\hat{\pi}\}_{i=1}^n$ of the moment conditions describing μ_0, in order to incorporate the extra information $\mu_0 = \mathbb{E}[q(x,\eta_0)]$ (see Section 3.3 for further details). That is,

$$\sum_{i=1}^{n} \hat{\pi}_i \frac{d_i}{\hat{p}(x_i)} \psi(z_i, \hat{\gamma}_{CLQ}) = 0.$$

Under correct specification, Chen et al. [7] show that $\hat{\gamma}_{CLQ}$ is locally efficient (see Section 3.2.1), and that $\hat{\gamma}_{CLQ}$ guarantees an efficiency gain over the simple Horvitz and Thompson [21] estimator even when $q(x,\eta_0) \neq \mathbb{E}[\psi(z,\gamma_0)|x]$. On the other hand, the class of augmented IPW estimators described in Section 2.3.1 may be more unstable in the sense that they do not guarantee such a property. This is despite augmented IPW estimation being based on doubly robust estimating equations and consequently the consistency of such estimators do not rely on correctly specifying $\mathbb{E}[\psi(z,\gamma_0)|x]$. For example, see Qin et al. ([31], p. 1500) and Ibrahim et al. ([22], p.340) for simulation evidence of the poor performance of augmented IPW estimators when a large proportion of observations on y are missing and $p_0(x)$ is highly dependent on x.

3 IPT Estimation with Generated Regressors

This section studies the properties of IPT estimators when further nuisance parameters enter the potentially overfitted propensity score. There are two ways in which estimation of a (possibily infinite-dimensional) nuisance parameter η_0 may enter the approximating functions $t(x, \eta_0)$ used in the propensity score estimating equation (2.8).

First the true propensity score in Assumption 2.5 is modelled by the more general specification $p_0(x) = G(r(x, \eta_0)'\delta_0)$. Secondly the assumption on the conditional expectation function (2.6) is weakened to $\mathbb{E}[\psi(z, \gamma_0)|x] = \Pi_0^\star t^\star(x, \eta_0)$. Those assumptions are re-stated accordingly.

Assumption 3.1 (Propensity score model). *There is a unique $\delta_0 \in int(\mathcal{D})$ where $\mathcal{D} \subset \mathbb{R}^{d_r}$ is compact and, for some $\epsilon > 0$, for any $||\eta - \eta_0|| < \epsilon$, $r(x, \eta)$ a d_r-dimensional known vector of linearly independent functions of x such that $p_0(x) = G(r(x, \eta_0)'\delta_0)$ for all $x \in \mathcal{X}$, where G is a known function such that* **(i)** *$G(.)$ is strictly increasing and continuously differentiable,* **(ii)** *$\lim_{v\to-\infty} G(v) = 0$ and $\lim_{v\to\infty} G(v) = 1$,* **(iii)** *$0 < \kappa \leq G(r(x, \eta)'\delta) \leq 1$ for all $\delta \in \mathcal{D}$ and $x \in \mathcal{X}$, and* **(iv)** *$G_1(r(x, \eta)'\delta_0) \leq \kappa_1$ for all $x \in \mathcal{X}$, where $G_1(a) = \partial G(a)/\partial a$. Furthermore, for some $\epsilon > 0$, for any $||\eta - \eta_0|| < \epsilon$, there exist $b_r(x) \geq 0$ and $b_{\partial r}(x) \geq 0$ such that $\sup_{\eta\in\mathcal{N}} ||r(x, \eta)|| < b_r(x)$ and $\sup_{\eta\in\mathcal{N}} ||\partial r(x, \eta)/\partial\eta|| < b_{\partial r}(x)$ where $\mathbb{E}[b_r(x)] < \infty$ and $\mathbb{E}[b_{\partial r}(x)] < \infty$.*

Assumption 3.2 (Conditional expectation function). *There is a unique $d_\psi \times d_{t^\star}$ matrix, $\Pi_0^\star$ and a $d_{t^\star}$-vector of linearly independent known functions $t^\star(x, \eta)$, continuous in the unknown parameter η and observables $x \in \mathcal{X}$, with a constant as first element, such that* **(i)** *there exist $b_{t^\star}(x) \geq 0$ and $b_{\partial t^\star}(x) \geq 0$ such that $||t^\star(x.\eta)|| \leq b_{t^\star}(x)$ with $\mathbb{E}[b_{t^\star}(x)] < \infty$, $||\partial t^\star(x, \eta)/\partial\eta|| \leq b_{\partial t^\star}(x)$ with $\mathbb{E}[b_{\partial t^\star}(x)] < \infty$, and* **(ii)**

$$\mathbb{E}[\psi(z, \gamma_0)|x] = \Pi_0^\star t^\star(x, \eta_0).$$

Furthermore, for $b_t(x) = \max\{b_r(x), b_{t^\star}(x)\}$ and $b_{\partial t}(x) = \max\{b_{\partial r}(x), b_{\partial t^\star}(x)\}$, $\mathbb{E}[b_t(x)^2] < \infty$, $\mathbb{E}[b_t(x)b_{\partial t}(x)] < \infty$ and, for $b_\psi(z)$ defined in Assumption 2.1, $\mathbb{E}[b_\psi(z)b_t(x)] < \infty$ and $\mathbb{E}[b_\psi(z)b_{\partial t}(x)] < \infty$.

Assumptions 3.1 and 3.2 also contain some boundedness conditions that allow for the application of uniform law of large numbers (for example, Lemma 2.4 of Newey and McFadden [27]) arguments for derivations. The motivation for this generalised set-up is now discussed.

3.1 Motivation for generated regressors

Justifying the MAR assumption for observational level data is not always straight-forward. Perhaps only after accounting for many, highly-nonlinear functions of x can d be regarded of as truly independent of y. With the greater availability of rich covariate data, estimation of the

propensity score based on machine learning techniques has been proposed, see, for example, Belloni et al. [4].

Assumption 3.1 permits a semiparametric approach to propensity score estimation where η_0 can be considered a generated regressor. Three-step treatment effects estimators that include generated regressors in the first step are studied by Heckman et al. [16], and Hahn and Ridder ([14], Section 4). Such sequential estimation procedures are popular in practice, and a generated regressors approach may also help to reduce the dimension of approximating functions needed.

In particular, note from (2.7) and (2.8) that the dimension of the IPT estimation problem increases one for one with the dimension of δ_1. The incorporation of generated regressors η_0 may reduce the dimension of the estimation problem by using effectively summarised covariate information.

If η_0 is nonparametrically estimated, the curse of dimensionality and general finite sample performance may ordinarily be of concern. However, doubly robust estimating equations permit larger smoothing biases resulting from nonparametric estimation (Firpo and Rothe [9]). Furthermore, via simulation, Frolich et al. [10] illustrate that the mean square error of doubly robust estimators is insensitive to the choice of tuning parameters for nuisance estimation. The simulation results in Section 4 further illustrate this robustness property.

As noted in Section 2.3.1, estimation of $\mathbb{E}[\psi(z,\gamma_0)|x]$ is necessary to implement any estimator based on the efficient influence function. Unlike propensity score estimation, nonparametric estimation of $\mathbb{E}[\psi(z,\gamma_0)|x]$ can be considerably more challenging, especially when $\mathbb{E}[\psi(z,\gamma_0)|x]$ is non-linear in functions of x. Following Robins et al. [34], regression models are often specified to estimate this function. For example, Wang et al. [41]'s partially linear model for $\mathbb{E}[\psi(z,\gamma_0)|x]$ for estimation of population means with missing data satisfies Assumption 3.2.

At the same time, by setting $\eta_0(x) = \mathbb{E}[y|x]$, nonparametric regressions included in the construction of approximating functions $t(x,\eta_0)$ allows IPT methods to directly exploit correlations with the conditional expectation function $\mathbb{E}[\psi(z,\gamma_0)|x]$. For example, suppose some of the missing data y concerns a binary variable y_1. Then estimating $\mathbb{E}[\psi(z,\gamma_0)|x]$ may involve estimating $\eta_0(x) = \mathbb{E}[y_1|x]$. As in Section 6.2 of GPE researchers may model the probability $\mathbb{P}(y_1 = 1|x)$ as $F(x,\eta_0)$ for some scalar function $F(.)$. By the MAR assumption, maximum likelihood estimation of η_0 is asymptotically valid. $F(x,\hat{\eta})$ would therefore contain information that should be included in the overfitted propensity score, and this is permitted by Assumption 3.2.

Finally, we note the leeway that the double robustness property offers may also explain why parametric and semiparametric specifications are popular in applications. It is likely that researchers have more confidence in specifying restrictive nuisance functions since one model being misspecified is not detrimental to consistent estimation of the parameter of interest.

3.2 IPT estimation with generated regressors

The estimator IPTGR studied here is the IPT estimator described in Section 2.3.3, but with plug-in estimates of η_0 entering $r(x,\eta_0)$ defined in Assumption 3.1 or entering $t^\star(x,\eta_0)$ of Assumption 3.2.

The IPTGR estimator of δ_0 is the solution $\hat{\delta}$ to (3.2) and the IPTGR estimator of γ_0 is the solution $\hat{\gamma}$ to (3.1) where

$$\frac{1}{n}\sum_{i=1}^{n}\frac{d_i\psi(z_i,\hat{\gamma})}{G(t(x_i,\hat{\eta})'\hat{\delta})}=0 \tag{3.1}$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\Big(\frac{d_i}{G(t(x_i,\hat{\eta})'\hat{\delta})}-1\Big)t(x_i,\hat{\eta})=0. \tag{3.2}$$

Two possibilities are considered for η_0: one in which η_0 is the solution to an unconditional moment restriction and the other in which η_0 is a conditional expectation that can be estimated by nonparametric regression.

3.2.1 Auxiliary moment restriction

Suppose there exists a moment restriction that identifies a finite-dimensional parameter η_0 and that may also contain auxiliary information, for example, about population means, as in Hellerstein and Imbens [18] and Bravo [5], that is,

$$\mathbb{E}[g_\eta(w,\eta_0)]=0, \tag{3.3}$$

where $w=(d,y,x)$. The moment indicator $g_\eta(w,\eta)$ is of dimension greater than that of η_0 and describes some causal relationship between y and x that may be required for approximating $\mathbb{E}[\psi(z,\gamma_0)|x]$. In the example when y includes a binary variable y_1 discussed in Section 3.1, $g_\eta(w,\eta)$ can include the score equations of a log-likelihood used to estimate $\mathbb{P}(y_1=1|x)=F(x,\eta_0)$, that is, $g_\eta(w,\eta)=d\partial[y_1\log(F(x,\eta))+(1-y_1)\log(1-F(x,\eta))]/\partial\eta$ (cf. GPE, p.1072). By MAR, (3.3) is then a valid moment condition.

The usual GMM or GEL procedures may be used to estimate η_0 under the usual regularity conditions. It is assumed that the standard assumptions for asymptotic normality of GEL estimation of η_0 hold, for example, Assumptions 1 and 2 of Newey and Smith [28], p.226. Let $\hat{\eta}$ denote a GEL estimator for η_0.

To simplify notation, define the following quantities: $\psi=\psi(z,\gamma_0)$, $t=t(x,\eta_0)$, $g_\eta=g(w,\eta_0)$, $\Psi=\mathbb{E}[\partial\psi(z,\gamma_0)/\partial\gamma]$, $G_\eta=\mathbb{E}[\partial g_\eta(w,\eta_0)/\partial\eta]$, $G=G(t(x,\eta_0)'\delta_1)$, $G_1(a)=\partial G(a)/\partial a$, $G_1=G_1(t(x,\eta_0)'\delta_1)$, $\partial t/\partial\eta=\partial t(x,\eta_0)/\partial\eta$, $\Omega_\eta=\mathbb{E}[g_\eta(w,\eta_0)g_\eta(w,\eta_0)']$ and $q_0(x,\gamma_0)=\mathbb{E}[\psi(z,\gamma_0)|x]$.

The local efficiency property considered here is now defined.

Definition (***local efficiency***)**.** An estimator $\hat{\gamma}$ of γ_0 is locally efficient if it has an asymptotic variance at most equal to

$$\Psi^{-1}\mathbb{E}\Big[\frac{\text{Var}(\psi(z,\gamma_0)|x)}{G} + q_0(x,\gamma_0)q_0(x,\gamma_0)'\Big]\Psi'^{-1}.$$

This is the semiparametric efficiency lower bound if estimation of γ_0 is based on Assumptions 2.1-2.4 and 3.1 (see Graham [11], p.439). Since restrictions on the propensity score do not lead to an efficiency gain, Assumption 3.1 does not allow for the possibility for a lower asymptotic variance of a regular estimator of γ_0. However, for the local efficiency property, neither the information contained in Assumption 3.2 nor the moment restriction (3.3) features. Accounting for such information on $q_0(x,\gamma_0)$ is likely to lead to the possibility of a lower asymptotic variance.

The following result describes the efficiency and robustness properties for the IPTGR estimator of γ_0 that solves (3.1) and (3.2).

Proposition 3.1 (IPTGR Estimation with plug-in $\hat{\eta}$). *Under Assumptions 2.1-2.4 and 3.1-3.2, and under usual assumptions required for GEL estimation of* (3.3)*, the IPTGR estimator $\hat{\gamma}$ that solves* (3.1) *and* (3.2)

(a) *is consistent*;

(b) *is locally efficient*;

(c) *has the following variance structure if Assumption 3.2 does not hold*

$$\sqrt{n}(\hat{\gamma}-\gamma_0) \xrightarrow{d} \mathcal{N}\big(0, \Psi^{-1}(\mathcal{V}_{(1)} - \mathcal{V}_{(2)} + \mathcal{V}_{(3)})\Psi'^{-1}\big),$$

where $\Sigma_1 = \mathcal{V}_{(1)} - \mathcal{V}_{(2)} + \mathcal{V}_{(3)}$ *with* $\mathcal{V}_{(1)} = \mathbb{E}\Big[\frac{\psi\psi'}{G}\Big]$, $\mathcal{V}_{(2)} = \mathbb{E}[\psi t']\mathbb{E}\Big[\frac{G}{1-G}tt'\Big]^{-1}\mathbb{E}[\psi t']'$ *and* $\mathcal{V}_{(3)} = CG_\eta^{-1}B_\eta' + B_\eta G_\eta^{-1}C' + CG_\eta^{-1}\Omega_\eta G_\eta^{-1}C'$. *Here* $B_\eta = \mathbb{E}[\psi g_\eta'] - \mathbb{E}\Big[\frac{G_1}{G}\mathbb{E}[\psi|x]t'\Big]\mathbb{E}\Big[\frac{G_1}{G}tt'\Big]^{-1}$ $\times\mathbb{E}\Big[\big(\frac{d}{G}-1\big)tg_\eta'\Big]$ *, and* $C = \mathbb{E}\Big[\frac{G_1}{G}\mathbb{E}[\psi|x]\delta_1'\frac{\partial t}{\partial\eta}\Big] - \mathbb{E}\Big[\frac{G_1}{G}\mathbb{E}[\psi|x]t'\Big]\mathbb{E}\Big[\frac{G_1}{G}tt'\Big]^{-1}\mathbb{E}\Big[\frac{G_1}{G}t\delta_1'\frac{\partial t}{\partial\eta}\Big]$;

(d) *is doubly robust: if at least one of Assumptions 3.1 and 3.2, and Assumptions 2.1-2.4 hold, $\hat{\gamma}$ is consistent.*

Remark 3.1(i): local efficiency. Part (b) of Proposition 3.1 states that under correct specification, there is no loss in asymptotic efficiency as compared with the original IPT estimator discussed in Section 2.3.3. For the IPT estimating equations (3.1) and (3.2), since η_0 enters through the propensity score model, the insensitivity of the asymptotic variance structure to propensity score estimation implies insensitivity to the estimation of η_0. This result is also related to the findings Statements 3 and 8 of Theorem 2.2 of Prokhorov and Schmidt ([30], p.49). In particular, when certain derivatives of the moment function with respect to nuisance parameters are zero, and when the model is just-identified, nuisance esti-

mation leaves the asymptotic variance unchanged, and one-step and two-step procedures have equivalent first-order properties under correct specification.

Remark 3.1(ii): variance structure when Assumption 3.2 does not hold. If Assumption 3.2 does not hold, the matrix C that appears in $\mathcal{V}_{(3)}$ is non-zero. On one hand, the information contained in the moment restriction (3.3) is incorporated which should lead to more efficient estimation. However, there is also an extra estimation error contributed by estimation of η_0. The original IPT estimator of Section 2.3.3 has an asymptotic variance structure of $\mathcal{V}_{(1)} - V_{(2)}$. Therefore, it is not immediately apparent whether estimated generated regressors in the IPT estimating equations yields an efficiency improvement if Assumption 3.2 does not hold.

Remark 3.1(iii): double robustness. Part (d) of Proposition 3.1 states that a double robustness property is retained. If at least one of the propensity score or the conditional expectation function is correctly specified, the estimator that solves (3.1) and (3.2) is consistent. Note that misspecification of the conditional expectation function can occur if $\mathbb{E}[\psi(z,\gamma_0)|x] \neq \Pi_0^\star t^\star(x,\eta)$ for any $\eta \in \mathcal{N}$ (where $\mathcal{N}$ is a compact set containing η_0 in its interior), and/or η_0 is not correctly described by the moment restriction (3.3). Therefore misspecification of the moment condition restriction (3.3) is unharmful for consistent estimation of γ_0 as long as η_0 does not enter a correctly specified model of the propensity score.

Similar properties hold if η_0 is an unknown conditional expectation that is nonparametrically estimated.

3.2.2 Nonparametric estimation of η_0

Suppose now that $\eta_0 := \eta_0(x_2)$ is the conditional expectation $\mathbb{E}[z_2|x_2]$, where z_2 is a subset of z and x_2 a subset of x. Even if z_2 contains y, the MAR assumption $d \perp y|x$ implies that $\eta_0(x_2)$ is identified. Let $\hat{\eta}(x_2)$ be any nonparametric estimator of $\eta_0(x_2)$ such that $|\hat{\eta}(x_2) - \eta_0(x_2)| = o_p(n^{-\frac{1}{4}})$; kernel, series etc. methods under the usual assumptions fulfil this requirement if η_0 is sufficiently smooth. The following efficiency and robustness properties for the IPTGR estimator of γ_0 that solves (3.1) and (3.2) with plug-in estimate $\hat{\eta}(x_2)$ hold.

Proposition 3.2 (IPTGR Estimation with nonparametric plug-in $\hat{\eta}(x_2)$). *Under Assumptions 2.1-2.4, 3.1-3.2, and Assumptions C.1, C.2 and C.3 of Appendix C, the IPTGR estimator $\hat{\gamma}$ that solves* (3.1) *and* (3.2)

(a) *is consistent*;

(b) *is locally efficient if $x_2 = x$, that is, the nonparametric regression averages over all observables x*;

(c) *has the following variance structure if Assumption 3.2 does not hold*

$$\sqrt{n}(\hat{\gamma}-\gamma_0) \xrightarrow{p} \mathcal{N}\big(0, \Psi^{-1}(\mathcal{V}_{(1)}-\mathcal{V}_{(2)}+\mathcal{V}_{(NP)})\Psi'^{-1}\big),$$

where $\Sigma_{NP} = \mathcal{V}_{(1)} - \mathcal{V}_{(2)} + \mathcal{V}_{(NP)}$ *with* $\mathcal{V}_{(1)}$ *and* $\mathcal{V}_{(2)}$ *are as described in Proposition 3.1, and* $\mathcal{V}_{(NP)} = \mathbb{E}[v(x_2)v(x_2)'(z_2-\eta_0(x_2))^2] - \Big\{\mathbb{E}[\psi v(x_2)'(z_2-\eta_0(x_2))] - EF^{-1}\mathbb{E}\Big[tv(x_2)'\Big(\frac{d}{G}-1\Big)(z_2-\eta_0(x_2))\Big]\Big\} - \Big\{\mathbb{E}[\psi v(x_2)'(z_2-\eta_0(x_2))] - EF^{-1}\mathbb{E}\Big[tv(x_2)'\Big(\frac{d}{G}-1\Big)(z_2-\eta_0(x_2))\Big]\Big\}'$. *Here* $v(x_2) = \mathbb{E}\Big[\frac{G_1}{G}\psi\delta_1'\frac{\partial t}{\partial \eta}\Big|x_2\Big] - EF^{-1}\mathbb{E}\Big[\frac{G_1}{G}t\delta_1'\frac{\partial t}{\partial \eta}\Big|x_2\Big]$, $E = \mathbb{E}\Big[\frac{G_1}{G}\psi t'\Big]$, $F = \mathbb{E}\Big[\frac{G_1}{G}tt'\Big]$;

(d) *is doubly robust: if at least one of Assumptions 3.1 and 3.2, and Assumptions 2.1-2.4 hold,* $\hat{\gamma}$ *is consistent.*

Remark 3.2: local efficiency. The proof of part (b) of Proposition 3.2 relies on the nonparametric regression averaging over all x, that is, $\eta_0(x) = \mathbb{E}[z_2|x]$, allowing the matrix $v(x_2)$ to be zero, which removes the estimation effect of η_0 from the variance calculations. This suggests attaining local efficiency may be difficult when the dimension of x is high due to the usual issues relating to the curse of dimensionality.

The following discussion shows that a tension exists between efficient use of auxiliary information and IPT robustness properties.

3.3 Using auxiliary moment information

If η_0 is just-identified by the unconditional moment restriction (3.3), then a plug-in estimator $\hat{\eta}$ into (3.1) and (3.2) is as efficient as one-step GMM or GEL estimation that stacks all available moment conditions $\mathbb{E}[d\psi(z,\gamma_0)/G(t(x,\eta_0)'\delta_1)] = 0$, $\mathbb{E}[((d/G(t(x,\eta_0)'\delta_1)-1)t(x,\eta_0)] = 0$ and $\mathbb{E}[g_\eta(w,\eta_0)] = 0$ together (see Remark 3.1(i)).

When η_0 is over-identified or if $g_\eta(w,\eta)$ contains other known population moment conditions, then an estimator of γ_0 that solves estimating equations (3.1) and (3.2) with plug-in estiamtor $\hat{\eta}$ may no longer be efficient. Under correct specification, one-step estimation of all moment conditions stacked together remains efficient (see, for example, Qin et al. [31]). However, some appealing robustness properties may not hold, as discussed in Section 3.4 below.

In order to guarantee efficiency gains from using moment condition (3.3) if Assumption 3.2 fails, consider re-weighting IPT estimating equations by weights obtained by GEL estimation that incorporates information from the auxiliary moment restriction (3.3). Such an estimator preserves the double robustness properties of the original IPT estimator as long as (3.3) holds.

3.3.1 GEL implied probabilities

Let d_{g_η} be the dimension of the moment indicator $g_\eta(w,\eta)$, and $d_\eta < d_{g_\eta}$ be the dimension of the parameter vector η. Given a sample $\{w_i\}_{i=1}^n$, η_0 is estimated by GEL as follows. Let

$$\hat{P}_n(\eta,\lambda) = \frac{1}{n}\sum_{i=1}^{n}[\rho(\lambda' g_\eta(w_i,\eta)) - \rho_0]$$

where the function $\rho(.)$ is concave on its domain $\mathcal{V}$, an open interval containing zero, with derivatives $\rho_j(v) = \partial^j \rho(v)/dv^j$, $\rho_j(0) = \rho_j$, $j = 0,1,...$, normalised without loss of generality as $\rho_1 = \rho_2 = -1$. The GEL estimator (Smith [39]) of η_0 is defined as

$$\hat{\eta} = \underset{\eta\in\mathcal{N}}{\operatorname{argmin}} \sup_{\lambda\in\hat{\Lambda}_n(\eta)} \hat{P}_n(\eta,\lambda)$$

for $\hat{\Lambda}_n(\eta) = \{\lambda : \lambda' g_\eta(w_i,\eta) \in \mathcal{V}, w_i \in \mathcal{W}, i = 1,...,n\}$. For any $\eta \in \mathcal{N}$, an estimator of the d_{g_η}-vector of auxiliary parameters λ is given by $\hat{\lambda}(\eta) = \operatorname{argmax}_{\eta\in\mathcal{N}} \hat{P}_n(\eta,\lambda)$. The first-order condition for λ imposes the sample moment constraint $\sum_{i=1}^n \hat{\pi}_i g_\eta(w_i,\hat{\eta},\hat{\lambda}) = 0$, where $\hat{\lambda} = \hat{\lambda}(\hat{\eta})$, and the GEL implied probabilities are

$$\hat{\pi}_i = \frac{\rho_1\big(\hat{\lambda}' g_\eta(w_i,\hat{\eta},\hat{\lambda})\big)}{\sum_{j=1}^n \rho_1\big(\hat{\lambda}' g_\eta(w_j,\hat{\eta},\hat{\lambda})\big)}, \quad (i = 1,...,n).$$

The vector λ has the interpretation of being the Lagrange multiplier associated with the sample moment constraint $\sum_{i=1}^n \hat{\pi}_i g_\eta(w_i,\hat{\eta},\hat{\lambda}) = 0$. Special cases include:

- empirical likelihood: $\rho(v) = \ln(1-v)$ and $\mathcal{V} = (-\infty, 1)$, resulting in implied probabilities $\hat{\pi}_i^{EL} = n^{-1}\big(1 + \hat{\lambda}' g_\eta(w_i,\hat{\eta},\hat{\lambda})\big)^{-1}$ $(i = 1,...,n)$.
- exponential tilting: $\rho(v) = -\exp(v)$, resulting in implied probabilities $\hat{\pi}_i^{ET} = \exp\big(\hat{\lambda}' g_\eta(w_i,\hat{\eta},\hat{\lambda})\big) / \sum_{j=1}^n \exp\big(\hat{\lambda}' g_\eta(w_i,\hat{\eta},\hat{\lambda})\big)$ $(i = 1,...,n)$.

3.3.2 GEL-weighted IPT estimation

Let $\{\hat{\pi}_i\}_{i=1}^n$ be the implied probabilities from GEL estimation of the auxiliary moment restriction (3.3). A GEL-weighted IPT estimator solves for $(\hat{\gamma},\hat{\delta})$ as

$$\sum_{i=1}^{n} \hat{\pi}_i \frac{d_i \psi(z_i,\hat{\gamma})}{G(t(x_i,\hat{\eta})'\hat{\delta})} = 0 \tag{3.4}$$

$$\sum_{i=1}^{n} \hat{\pi}_i \Big(\frac{d_i}{G(t(x_i,\hat{\eta})'\hat{\delta})} - 1\Big) t(x_i,\hat{\eta}) = 0. \tag{3.5}$$

That is, the EDF $1/n$ weights for method of moments estimation, (3.1) and (3.2), have been replaced by the GEL implied probability weights that contain information from the auxiliary moment restriction (3.3).

This method is also operational in the case where (3.3) describes a relationship for complete case units only. For example, as described in Section 3.2.1, if y is a binary variable and its conditional distribution function is modelled by $\mathbb{P}(y = 1|x) = F(x, \eta_0)$, then the moment functions based on the maximum likelihood score equations are $g_\eta(w, \eta) = d \times \partial[y_1 \log(F(x, \eta)) + (1 - y_1) \log(1 - F(x, \eta))]/\partial\eta$. Since for $d = 0$, $g_\eta(w, \eta) = 0$ for any η, the GEL implied probabilities satisfy $\hat{\pi}_i = \rho(0)/\sum_{i=1}^{n} \rho(0) = 1/n$ for all observations $d_i = 0$, $(i = 1, ..., n)$. That is, the implied probabilities are uninformative for the sample $d_i = 0$, $(i = 1, ..., n)$, however they are informative as usual for observations $d_i = 1$, $(i = 1, ..., n)$.

Proposition 3.3 (IPT estimation with GEL implied probability weights). *Under Assumptions 2.1-2.4 and 3.1-3.2, and under usual assumptions required for GMM/GEL estimation of* (3.3)*, the IPT estimator* $\hat{\gamma}$ *that solves* (3.4) *and* (3.5)

(a) *is consistent;*

(b) *is more efficient than the IPT estimator based on* (3.1) *and* (3.2)*;*

(c) *is more efficient than the IPT estimator based on* (3.1) *and* (3.2) *if Assumption 3.2 does not hold;*

(d) *is doubly robust: if at least one of Assumptions 3.1 and 3.2, and Assumptions 2.1-2.4 hold,* $\hat{\gamma}$ *is consistent.*

Remark 3.3(i): efficiency. Part (b) of Proposition 3.3 states that under correct specification, the GEL-weighted IPT estimator is locally efficient. Furthermore, the variance matrix is reduced further than the locally efficient variance matrix stated in Section 3.2.1 by a positive definite matrix. The extent of the variance reduction depends on a full rank $d_\psi \times d_{g_\eta}$ matrix $B_\eta = \mathbb{E}[\psi g_\eta'] - \mathbb{E}\Big[\frac{G_1}{G}\mathbb{E}[\psi|x]t'\Big]\mathbb{E}\Big[\frac{G_1}{G}tt'\Big]^{-1}\mathbb{E}\Big[\big(\frac{d}{G} - 1\big)tg_\eta'\Big]$, which is the correlation between the auxiliary moment function $g_\eta(w, \eta_0)$ and a linear combination of moment functions that describe γ_0 and δ_0. The greater the correlation, the greater the gain in efficiency. An analogous statement can be made for part (c) of Proposition 3.3.

Remark 3.3(ii): double robustness i. Part (d) of Proposition 3.3 suggests that if at least one of the propensity score or the conditional expectation function is correctly specified, then the IPT estimator that solves (3.4) and (3.5) remains consistent. Use of the GEL weights $\hat{\pi}_i$ $(i = 1, ..., n)$ therefore preserves a double robustness property with the estimating equation for γ_0 being an appropriate linear combination of the sample moment conditions required for double robustness.

Remark 3.3(iii): double robustness ii. In contrast to Remark 3.1(iii) for an IPT estimator not weighted by $\hat{\pi}_i$ $(i = 1, ..., n)$, the validity of the moment condition (3.3) is required

for the above results to hold. Therefore, if the researcher is not confident on the specification of the auxiliary moment restriction (3.3), then the unweighted IPT estimator with plug-in $\hat{\eta}$ is a more sensible estimation method for γ_0.

REMARK 3.3(IV): GEL-WEIGHTING. In Chapter 2 of the thesis, the theoretical and finite sample properties of a similar GEL-weighted estimator are studied in a more general setting; therefore the simulation study in Section 4 does not consider the use of auxiliary information and GEL implied probabilities.

3.4 One-step efficient estimation

Under the missing data set-up considered in this paper, if η_0 is described by (3.3), all the moment conditions available are

$$\mathbb{E}\Big[\frac{d\psi(z,\gamma_0)}{G(t(x,\eta_0)'\delta_1)}\Big] = 0 \tag{3.6}$$

$$\mathbb{E}\Big[\Big(\frac{d}{G(t(x,\eta_0)'\delta_1)} - 1\Big)t(x,\eta_0)\Big] = 0 \tag{3.7}$$

$$\mathbb{E}[g_\eta(w,\eta_0)] = 0 \tag{3.8}$$

When η_0 is over-identified, one-step GMM or GEL methods that are based on all moment conditions above are generally more efficient than two-step methods. One-step methods take the optimal linear combination of moment conditions for purposes of efficiency. However, such linear combinations may result in poor properties under misspecification.

Under misspecification of the propensity score model, suppose $\delta_\star$ is the psuedo-true value that satisfies the population moment function in (3.7). From Section 3.3.1, the following sample condition follows from GEL estimation[1]

$$\sum_{i=1}^{n} \hat{\pi}_i \frac{d_i}{G(t(x_i,\hat{\eta})'\hat{\delta})} t(x_i,\hat{\eta}) = \sum_{i=1}^{n} \hat{\pi}_i t(x_i,\hat{\eta}) \tag{3.9}$$

where the GEL implied probabilities $\{\hat{\pi}_i\}_{i=1}^n$ are based on all population moments (3.6), (3.7) and (3.8). A double robustness result relies on the population condition

$$\mathbb{E}\Big[\frac{\mathbb{E}[d|x]}{G(t(x,\eta_0)'\delta_\star)} t(x,\eta_0)\Big] = \mathbb{E}[t(x,\eta_0)] \tag{3.10}$$

(see Proof of Condition A.2 of Proposition 3.1(c)). However the sample moment condition (3.9) no longer implies (3.10) holds asymptotically under misspecification since the behaviour of $\hat{\pi}_i$ $(i = 1, ..., n)$ is not guaranteed to be stable, in particular, $\hat{\pi}_i$ $(i = 1, ..., n)$ may not remain close to $1/n$, see Imbens et al. ([23], p. 337). This suggests that a double robustness property

[1] If γ_0 is estimated by GMM estimation, the same arguments apply with the implied probabilities for GMM estimation being given by Back and Brown [2].

will not hold for one-step estimation of (3.6), (3.7) and (3.8). That is, consistent estimation of γ_0 relies on correct specification of all moment conditions including that for the propensity score.

4 Simulation Study

This section illustrates the use of IPT estimation with generated regressors and considers its finite sample performance in a simulation study. We consider the estimation of population means when data are missing; Example 2.1 of GPE shows such models can be used to study the prevalence of HIV for a population in which the probability that an individual takes an HIV test is based on individual characteristics.

Let $x_1 \sim N(0,1)$ and $x_2 \sim N(0,1)$ be independent covariates, and y be an outcome variable determined by $y = \mathcal{S}_0 t^\star(x) + u$, where $\mathcal{S}_0 = (0.5, 0.5, 0.5, 0.5, 0.5, 0.5)'$ and $t^\star(x) = (1, x_1, x_1x_2, x^2, x_1^2x_2, x_2^3)'$, and $u \sim N(0,1)$ is an error term independent of x_1 and x_2. The parameter of interest is $\gamma_0 = \mathbb{E}[y]$ and the moment function is $\psi(z,\gamma) = y - \gamma$. Therefore, Assumption 3.2 is satisfied with $\Pi_0^{\star\prime} = (0.5 - \gamma_0, 0.5, 0.5, 0.5, 0.5, 0.5)'$.

Let d be a binary variable which equals 1 if y is observed, and 0 if y is missing. The propensity score model is given by $\mathbb{P}(d = 1|x) = G(r(x)'\delta_0)$, where for any $a > 0$, $G(a) = exp(a)/(1 + exp(a))$, $r(x) = (1, x_{1,}, x_2)'$ and $\delta_0 = (1, -0.35, -0.35)'$.

The following estimators of γ_0 are considered for comparison.

4.1 IPW, IPT and imputation estimators

Under Assumption 3.1 and 3.2, the optimal IPT estimator, denoted $\hat{\gamma}_{IPT}$, over-fits the propensity score with the vector of approximating functions $t_1^\star(x) = (1, x_1, x_2, x_1x_2, x_2^2, x_1^2x_2, x_2^3)'$. In practice, however, it may be difficult to choose the correct approximating function $t_1^\star(x)$. In a bid to capture complex functionals of x_1 and x_2 which hypothetically may be relevant to approximate a highly non-linear function $\mathbb{E}[y|x_1, x_2]$, it is likely that IPT estimation may involve many functionals that are irrelevant. Thus we consider an IPT estimator with a vector of approximating functions that includes some irrelevant elements $t_2^\star(x) = (1, x_1, x_2, x_1x_2, x_1^2, x_2^2, x_1^2x_2, x_1x_2^2, x_1^3, x_2^3)$ denoted by $\hat{\gamma}_{IPTO}$ (IPTO). Finally, we consider an IPT estimator with generated regressors (IPTGR) $\hat{\gamma}_{IPT-NP}$ obtained from using the vector of approximating functions $t_3^\star(x) = (1, x_1, x_2, \hat{\eta}(x))'$, where $\hat{\eta}(x)$ is a nonparametric (local linear) kernel regression estimator of $\mathbb{E}[y|x_1, x_2]$ using Hayfield and Racine's [15] "np" R package, and where the bandwidths are computed by least squares cross validation.

Let $\hat{p}(x) = G(r(x)'\hat{\delta})$ be the estimated propensity score from a logistic regression of d on $(1, x_1, x_2)$. Then, the IPW estimator $\hat{\gamma}_{IPW}$ of γ_0 is given by $\hat{\gamma}_{IPW} = n^{-1}\sum_{i=1}^n d_i\hat{p}(x_i)^{-1}y_i$. The doubly robust estimator of Robins et al. [34] is given by $\hat{\gamma}_{DR} = (n^{-1}\sum_{i=1}^n d_i\hat{p}(x_i)^{-1}y_i) -$

$(n^{-1}\sum_{i=1}^{n}\hat{p}(x_i)^{-1}(d_i-\hat{p}(x_i))\hat{\eta}(x_i))$. Finally, the imputation estimator (IMP) of γ_0 is given by $\hat{\gamma}_{IMP}=n^{-1}\sum_{i=1}^{n}d_iy_i+(1+d_i)\hat{\eta}(x_i)$.

All the IPT estimators, along with IPW and IMP described above, are consistent under Assumptions 2.1-2.4, 3.1 and 3.2. In order to show the severity of the missing data problem, we also note the bias results of the crude, complete-case estimator of γ_0 that calculates the average of y using the observed values, that is, $\hat{\gamma}_{CC}=(\sum_{i=1}^{n}d_i)^{-1}\sum_{i=1}^{n}d_iy_i$.

4.2 Simulation results

Table 1 presents the averaged results from 1000 simulations. The standard errors are given by the square-root estimated asymptotic variance divided by $\sqrt{n}$. An estimate of the asymptotic variance of $\hat{\gamma}_{IPW}$ is given by $n^{-1}\sum_{i=1}^{n}((d_i\hat{p}(x_i)^{-1}y_i)-\hat{\gamma}_{IPW})^2$. An estimate of the asymptotic variance of $\hat{\gamma}_{DR}$ is $n^{-1}\sum_{i=1}^{n}((d_i\hat{p}(x_i)^{-1}y_i)-(\hat{p}(x_i)^{-1}(d_i-\hat{p}(x_i))\hat{\eta}(x_i))-\hat{\gamma}_{DR})^2$. A consistent estimator of the asymptotic variance of IPT estimators is given in GPE, equation (A.9), p.1075. The true value of γ_0 is given by $\gamma_0=\mathbb{E}[y]=0.5+0.5\mathbb{E}[x_2^2]=1$.

		$\hat{\gamma}_{CC}$	$\hat{\gamma}_{IPW}$	$\hat{\gamma}_{IMP}$	$\hat{\gamma}_{DR}$	$\hat{\gamma}_{IPT}$	$\hat{\gamma}_{IPTO}$	$\hat{\gamma}_{IPTGR}$
$n=250$	Estimate	0.7504	0.9965	0.9352	0.9525	0.9937	0.9945	0.9513
	Standard Error	-	0.2339	-	0.1708	0.1803	0.1811	0.1671
	Standard Deviation	0.1853	0.2178	0.1763	0.1777	0.1785	0.1788	0.1770
$n=500$	Estimate	0.7586	1.0047	0.9610	0.9764	1.0032	1.0033	0.9756
	Standard Error	-	0.1670	-	0.1246	0.1293	0.1293	0.1218
	Standard Deviation	0.1362	0.1529	0.1259	0.1262	0.1288	0.1290	0.1260
$n=750$	Estimate	0.7651	1.0110	0.9709	0.9846	1.0056	1.0057	0.9839
	Standard Error	-	0.1392	-	0.1030	0.1058	0.1058	0.1007
	Standard Deviation	0.1096	0.1301	0.1067	0.1076	0.1092	0.1092	0.1075

Table 1. Simulation results of IPW, IPT and imputation estimators.

Firstly, the results of the complete case estimator confirm that ignoring the missing data problem leads to severe biases, and thus IPW or imputation based methods are suitable under the MAR assumption.

In terms of finite sample bias, IPW and IPT estimators that do not involve generated regressors perform very well. This advantage over the DR and IPTGR estimators appears to diminish as the sample size grows. This could be due to DR and IPTGR being functions of nonparametric regressions that are likely to have larger biases in small samples.

In terms of variance, however, the DR and IPTGR estimators perform the best, with IPTGR invariably being the best by this measure. IPTO has a slightly higher variance than IPT in

small samples, which would be the expected impact of adding irrelevant functionals in the construction of $t_2^\star(x)$.

The IPTGR method of including a plug-in estimate of $\mathbb{E}[y|x_1, x_2]$ directly focuses $t_3^\star(x)$ to include only relevant information and serves as a dimension-reduction technique in this case. Interestingly, IPTGR offers considerable gains in terms of lower finite-sample variances.

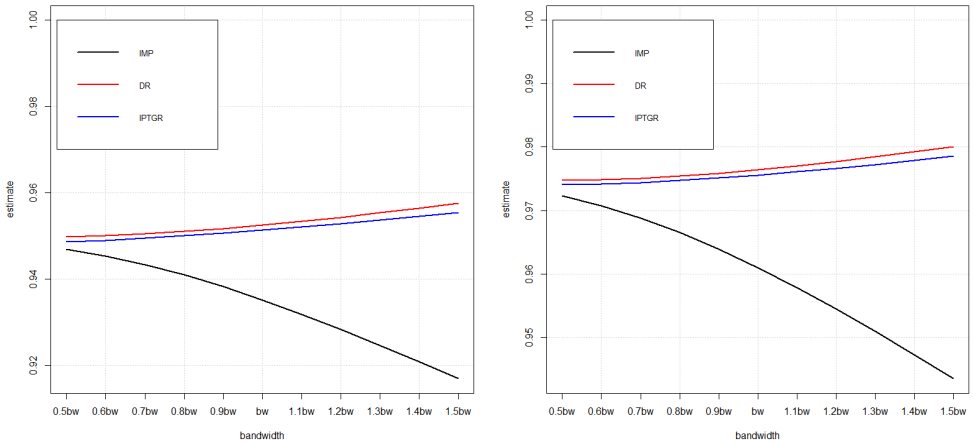
Overall, on the evidence of this simulation study, due to the incorporation of plug-in nonparametric estimates, IPTGR displays greater biases relative to IPW and IPT that may make its use undesirable in very small samples. However, in practice, correctly modelling $\mathbb{E}[y|x_1, x_2]$ may be dificult, and for larger sample sizes IPTGR performs very well.

4.3 IPTGR estimator is insensitive to bandwidth choice

DR, IPTGR and IMP are estimators of γ_0 that require with plug-in estimators $\hat{\eta}(x)$ of $\mathbb{E}[y|x]$. When $\hat{\eta}(x)$ is a local linear kernel regression of y on x_1 and x_2, the 2-dimensional optimal bandwidth bw is computed by least squares cross validation. Here, the sensitivity of bias and mean square error (MSE) properties of DR, IPTGR and IMP are compared when the estimators are computed over the range of bandwidth choices contained in $\{0.5\text{bw}, 0.6\text{bw}, ..., 1.4\text{bw}, 1.5\text{bw}\}$.

Figures 1 and 2 show that for small samples, the optimal bandwidth bw may under-smooth $\hat{\eta}(x)$ for the purpose of bias and MSE properties of DR and IPTGR. On the other hand, bw is over-smoothed for the bias and MSE properties of IMP.

The double robustness property suggests that even when a model for $\mathbb{E}[y|x_1, x_2]$ is misspecified, IPT estimators of γ_0 remain consistent as long as the propensity score model is correctly specified. In this case, when $\mathbb{E}[y|x]$ is nonparametrically estimated, the double robustness property suggests that DR and IPTGR should be insensitive to bandwidth choice (see, for example, Firpo and Rothe [9]) as compared with the imputation estimator IMP.



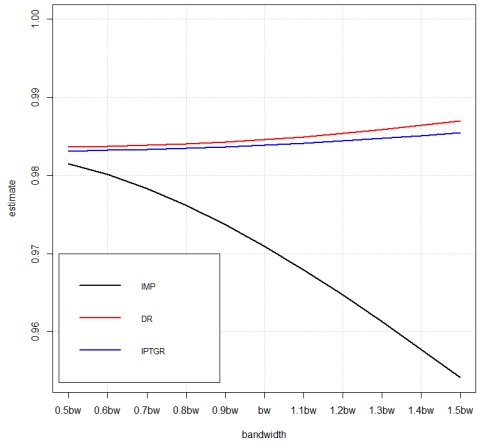


Figure 1. Sensitivity of estimators to bandwidth choice. bw is the optimal bandwidth as computed by least-squares cross validation. Top left: $n = 250$; top right: $n = 500$; bottom: $n = 750$.

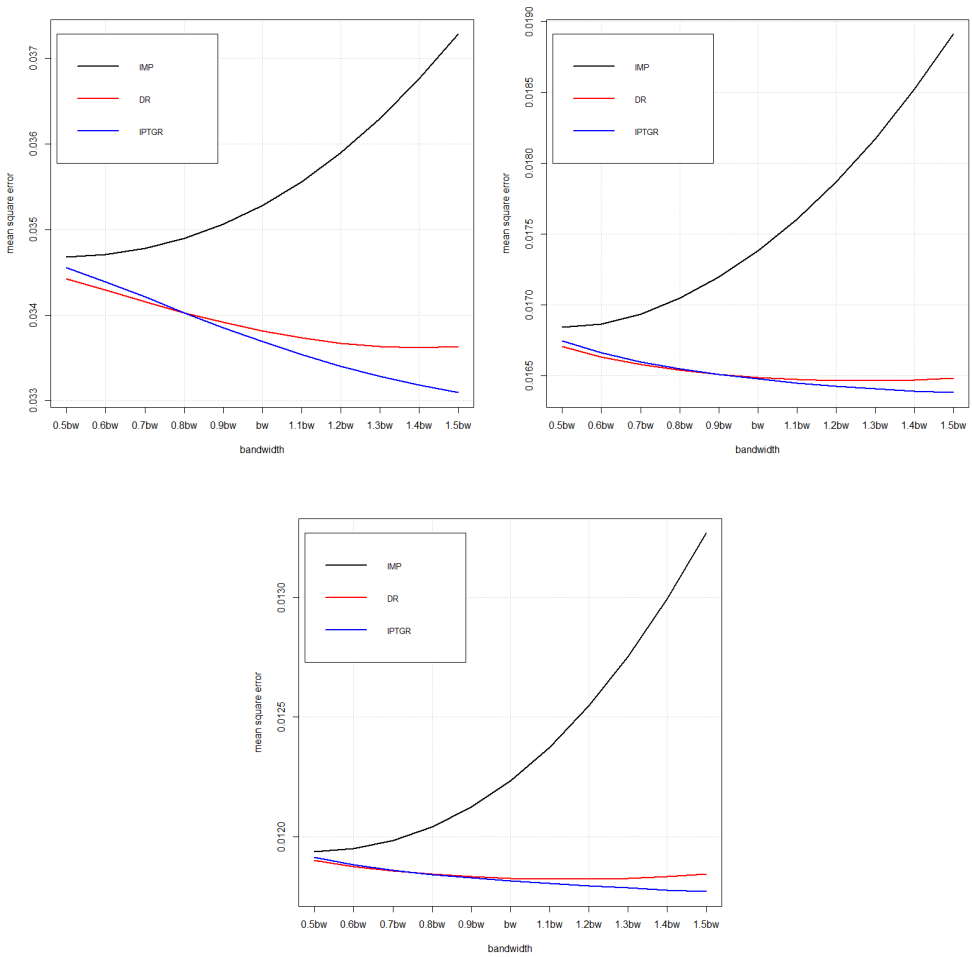


Figure 2. Sensitivity of MSE of estimators to bandwidth choice. bw is the optimal bandwidth as computed by least-squares cross validation. Top left: $n = 250$; top right: $n = 500$; bottom: $n = 750$.

Figures 1 and 2 confirm that DR and IPTGR are far less sensitive to the choice of bandwidth than IMP. Furthermore, the choice of bandwidth becomes even less influential as the sample size increases, that is, the bias and MSE of DR and IPTGR become flatter functions of the bandwidth.

5 Conclusion

This paper considers an extension of the IPT estimation method to allow semiparametric specifications of the propensity score and conditional expectation functions. It has been shown in Hirano et al. [20] and GPE that for the purpose of efficiency gain, it is useful to overfit the propensity score model to include functions that are correlated with the conditional expectation of the moment function given observables. Since this conditional expectation function is likely to involve unknown functions, it may be beneficial to include nonparametrically estimated functions as generated regressors in the propensity score. Local efficiency and double robustness properties are retained by this generalisation. Furthermore, if the generated regressors involve further nuisance estimation of unconditional moment restrictions, a method is proposed that maintains the double robustness result while allowing for an efficiency gain by using GEL implied probabilities.

While this method is also a useful way to reduce the dimensionality of a potentially overfitted propensity score model, it would be interesting to extend IPT methods to allow for high dimensional covariates. Since IPT estimation is designed to be insensitive to estimation errors arising from propensity score estimation, it may be the case that the estimation process also remains insensitive to propensity score estimation if variable selection methods are used to select covariates entering the propensity score model.

References

[1] Daniel Ackerberg, Xiaohong Chen, Jinyong Hahn, and Zhipeng Liao. Asymptotic efficiency of semiparametric two-step GMM. *Review of Economic Studies*, 2014.

[2] Kerry Back and David P. Brown. Implied Probabilities in GMM Estimators. *Econometrica*, 61(4):971–975, 1993.

[3] Heejung Bang and James M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–972, 2005.

[4] Alexandre Belloni, Victor Chernozhukov, I. Fernandez-Val, and Christian Hansen. Program Evaluation and Causal Inference With High-Dimensional Data. *Econometrica*, 85(1):233–298, 2017.

[5] Francesco Bravo. Efficient M-estimators with auxiliary information. *Journal of Statistical Planning and Inference*, 140(11):3326–3342, 2010.

[6] Gary Chamberlain. Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, 34(3):305–334, 1987.

[7] Song Xi Chen, Denis H Y Leung, and Jing Qin. Improving semiparametric estimation by using surrogate data. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 70(4):803–823, 2008.

[8] Xiaohong Chen, Han Hong, and Alessandro Tarozzi. Semiparametric efficiency in GMM models with auxiliary data. *Annals of Statistics*, 36(2):808–843, 2008.

[9] Sergio Firpo and Christoph Rothe. Properties of Doubly Robust Estimators when Nuisance Functions are Estimated Nonparametrically. 2016.

[10] Markus Frölich, Martin Huber, and Manuel Wiesenfarth. The finite sample performance of semi- and non-parametric estimators for treatment effects and policy evaluation. *Computational Statistics and Data Analysis*, 115:91–102, 2017.

[11] Bryan S. Graham. Efficiency Bounds for Missing Data Models With Semiparametric Restrictions. *Econometrica*, 79(2):437–452, 2011.

[12] Bryan S. Graham, Cristine Campos De Xavier Pinto, and Daniel Egel. Inverse probability tilting for moment condition models with missing data. *Review of Economic Studies*, 79(3):1053–1079, 2012.

[13] Jinyong Hahn. On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects. *Econometrica*, 66(2):315–331, 1998.

[14] Jinyong Hahn and Geert Ridder. Asymptotic Variance of Semiparametric Estimators With Generated Regressors. *Econometrica*, 81(1):315–340, 2013.

[15] Tristen Hayfield and Jeffrey S. Racine. Nonparametric Econometrics: The np Package. *Journal of Statistical Software*, 27(5), 2008.

[16] J. J. Heckman, H. Ichimura, and P. E. Todd. Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme. *The Review of Economic Studies*, 64(4):605–654, 1998.

[17] James J. Heckman. Sample Selection Bias as a Specification Error. *Econometrica*, 47(1):153–161, 1979.

[18] Judith K. Hellerstein and Guido W. Imbens. Imposing Moment Restrictions from Auxiliary Data by Weighting. *The Review of Economics and Statistics*, 81(1):1–14, 1999.

[19] Miguel A. Hernán, Emilie Lanoy, Dominique Costagliola, and James M. Robins. Comparison of dynamic treatment regimes via inverse probability weighting, 2006.

[20] Keisuke Hirano, Guido W. Imbens, and Ridder Geert. Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score. *Econometrica*, 71(4):1161–1189, 2003.

[21] D. G. Horvitz and D. J. Thompson. A Generalization of Sampling Without Replacement from a Finite Universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

[22] Joseph G Ibrahim, Ming-Hui Chen, Stuart R Lipsitz, and Amy H Herring. Missing-Data Methods for Generalized Linear Models. *Journal of the American Statistical Association*, 100(469):332–346, 2005.

[23] Guido W. Imbens, Richard H. Spady, and Phillip Johnson. Information theoretic approaches to inference in moment condition models. *Econometrica*, 66(2):333–357, 1998.

[24] Roderick J.A. Little and Donald B. Rubin. Statistical Analysis with Missing Data. *Wiley, New York.*, page 381, 1987.

[25] Whitney K. Newey. The Asymptotic Variance of Semiparametric Estimators. *Econometrica*, 62(6):1349–1382, 1994.

[26] Whitney K. Newey. Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, 79(1):147–168, 1997.

[27] Whitney K. Newey and Daniel McFadden. Large sample estimation and hypothesis testing. *Handbook of Econometrics*, 4:2111–2245, 1994.

[28] Whitney K. Newey and Richard J. Smith. Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica*, 72(1):219–255, 2004.

[29] Taisuke Otsu. Large deviations of generalized method of moments and empirical likelihood estimators. *The Econometrics Journal*, 14(2):321–329, 2011.

[30] Artem Prokhorov and Peter Schmidt. GMM redundancy results for general missing data problems. *Journal of Econometrics*, 151(1):47–55, 2009.

[31] Jing Qin, Biao Zhang, and Denis H. Y. Leung. Empirical Likelihood in Missing Data Problems. *Journal of the American Statistical Association*, 104(488):1492–1503, 2009.

[32] Joaquim J S Ramalho and Richard J. Smith. Goodness of Fit Tests for Moment Condition Models. 2006.

[33] James M. Robins. Non-response models for the analysis of non-monotone non-ignorable missing data. *Statistics in Medicine*, 16(1-3):21–37, 1997.

[34] James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of Regression Coefficients When Some Regressors Are Not Always Observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994.

[35] Donald B. Rubin. Inference and Missing Data. *Biometrika*, 63(3):581–592, 1976.

[36] Donald B Rubin. Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, 6(1):34–58, 1978.

[37] Donald B. Rubin. *Multiple imputation for nonresponse in surveys*. Wiley, 1987.

[38] Daniel O. Scharfstein, Andrea Rotnitzky, and James M. Robins. Adjusting for Nonignorable Drop-Out Using Semiparametric Nonresponse Models. *Journal of the American Statistical Association*, 94(448):1096–1120, 1999.

[39] Richard J. Smith. Alternative Semi-Parametric Likelihood Approaches to Generalised Method of Moments Estimation. *The Economic Journal*, 107(441):503–519, 1997.

[40] Anastasios Tsiatis. *Semiparametric Theory and Missing Data*. Springer Series in Statistics, 2006.

[41] Qihua Wang, Oliver Linton, and Wolfgang Härdle. Semiparametric Regression Analysis With Missing Response at Random. *Journal of the American Statistical Association*, 99(466):334–345, 2004.

[42] Qihua Wang and J. N K Rao. Empirical likelihood-based inference under imputation for missing response data. *Annals of Statistics*, 30(3):896–924, 2002.

[43] Jeffrey M. Wooldridge. Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics*, 141(2):1281–1301, 2007.

Appendix

Notation

As stated in Section 3.2, let $\psi = \psi(z, \gamma_0)$, $t = t(x, \eta_0)$, $g_\eta = g(w, \eta_0)$, $\Psi = \mathbb{E}[\partial\psi(z, \gamma_0)/\partial\gamma]$, $G_\eta = \mathbb{E}[\partial g_\eta(w, \eta_0)/\partial\eta]$, $G = G(t(x, \eta_0)'\delta_1)$, $G_1(a) = \partial G(a)/\partial a$, $G_1 = G_1(t(x, \eta_0)'\delta_1)$, $\partial t/\partial\eta = \partial t(x, \eta_0)/\partial\eta$, $\Omega_\eta = \mathbb{E}[g_\eta(w, \eta_0)g_\eta(w, \eta_0)']$, and $q_0(x; \gamma_0) = \mathbb{E}[\psi(z, \gamma_0)|x]$.

In addition, the following notation is maintained in the derivations. Let $\psi_i(\gamma) = \psi(z_i, \gamma)$, $\psi_i = \psi_i(\gamma_0)$, $\psi(\gamma) = \psi(z, \gamma)$, $\Psi_i = \partial\psi_i(\gamma_0)/\partial\gamma$, $\Psi_i(\gamma) = \partial\psi_i(\gamma)/\partial\gamma$, $\Psi(\gamma) = \partial\psi(z, \gamma)/\partial\gamma$, $t_i = t(x_i, \eta_0)$, $t(\eta) = t(x, \eta)$, $t_i(\eta) = t(x_i, \eta)$, $\hat{t}_i = t(x_i, \hat{\eta})$, $\partial t_i/\partial\eta = \partial t(x_i, \eta_0)/\partial\eta$, $\partial t_i(\eta)/\partial\eta = \partial t(x_i, \eta)/\partial\eta$, $g_{\eta i} = g(w_i, \eta_0)$, $G_i = G(r(x_i, \eta_0)'\delta_0)$, $\hat{G}_i = G(t(x_i, \hat{\eta})'\hat{\delta})$, $G_i(\eta) = G(t_i(\eta)'\delta_1)$, $G_{1i} = G_1(r(x_i, \eta_0)'\delta_0)$, $\hat{G}_{1i} = G_1(t(x_i, \hat{\eta})'\hat{\delta})$, $G_{1i}(\eta) = G_1(t_i(\eta)'\delta_1)$, $G(\eta) = G(t(\eta)'\delta_0)$ and $G_1(\eta) = G_1(t(\eta)'\delta_0)$.

Also let $C = \mathbb{E}\Big[\frac{G_1}{G}\mathbb{E}[\psi|x]\delta_1'\frac{\partial t}{\partial\eta}\Big] - \mathbb{E}\Big[\frac{G_1}{G}\mathbb{E}[\psi|x]t'\Big]\mathbb{E}\Big[\frac{G_1}{G}tt'\Big]^{-1}\mathbb{E}\Big[\frac{G_1}{G}t\delta_1'\frac{\partial t}{\partial\eta}\Big]$, $E = \mathbb{E}\Big[\frac{G_1}{G}\psi t'\Big]$ and $F = \mathbb{E}\Big[\frac{G_1}{G}tt'\Big]$.

A Proofs of Main Results

Proof of Proposition 3.1

Part (a) - Consistency

The proof for consistency is given in Appendix B; the arguments of the proof hold $\hat{\pi}_i = n^{-1}$, $(i = 1, ..., n)$. □

Part (c) - Variance structure under misspecification

Consider the case where η_0 is identified by the moment condition $\mathbb{E}[g_\eta(w, \eta_0)] = 0$. Then under the usual regularity conditions (Theorem 3.1 of Newey and McFadden [27], p.2143), the following asymptotically linear form from method of moments estimation of η_0 holds

$$\hat{\eta} - \eta_0 = -G_\eta^{-1}\frac{1}{n}\sum_{i=1}^{n} g_\eta(w_i, \eta_0) + o_p(n^{-\frac{1}{2}}). \tag{A.1}$$

The IPT estimating equation for the propensity score is

$$\frac{1}{n}\sum_{i=1}^{n}\Big(\frac{d_i}{\hat{G}_i} - 1\Big)\hat{t}_i = 0$$

By a Taylor expansion around δ_1, for some $\dot{\delta}$ on the line segment joining $\hat{\delta}$ and δ_1,

$$\frac{1}{n}\sum_{i=1}^{n}\Big(\frac{d_i}{G(\hat{t}_i'\delta_1)} - 1\Big)\hat{t}_i - \Big(\frac{1}{n}\sum_{i=1}^{n}\frac{d_i G_1(\hat{t}_i'\dot{\delta})}{G(\hat{t}_i'\dot{\delta})^2}\hat{t}_i\hat{t}_i'\Big)(\hat{\delta} - \delta_1) = 0.$$

By T,

$$\Big|\Big|\frac{1}{n}\sum_{i=1}^{n}\frac{d_iG_1(\hat{t}_i'\dot{\delta})}{G(\hat{t}_i'\dot{\delta})^2}\hat{t}_i\hat{t}_i' - \mathbb{E}\Big[\frac{dG_1}{G^2}tt'\Big]\Big|\Big| \;\leq\; \Big|\Big|\frac{1}{n}\sum_{i=1}^{n}\frac{d_iG_1(\hat{t}_i'\dot{\delta})}{G(\hat{t}_i'\dot{\delta})}\hat{t}_i\hat{t}_i' - \mathbb{E}\Big[\frac{dG_1(t(\hat{\eta})'\dot{\delta})}{G(t(\hat{\eta})'\dot{\delta})^2}t(\hat{\eta})t(\hat{\eta})'\Big]\Big|\Big| + \Big|\Big|\mathbb{E}\Big[\frac{dG_1(t(\hat{\eta})'\dot{\delta})}{G(t(\hat{\eta})'\dot{\delta})^2}t(\hat{\eta})t(\hat{\eta})'\Big] - \mathbb{E}\Big[\frac{dG_1}{G^2}tt'\Big]\Big|\Big|. \quad (A.2)$$

By Assumption 3.1 and 3.2 and CS, for any $\eta \in \mathcal{N}$, $\delta \in \mathcal{D}$, $||dG_1(t(\eta)'\delta)t(\eta)t(\eta)'/G(t(\eta)'\delta)^2|| \leq ||t(\eta)||^2\kappa_1/\kappa^2 \leq b_t(x)\kappa_1/\kappa^2$, where the last term on the RHS is bounded above in expectation. Also, $G(t(\eta)'\delta)$, $G_1(t(\eta)'\delta)$ and $t(\eta)$ are continuous in parameters $\eta \in \mathcal{N}$ and $\delta \in \mathcal{D}$, and $\mathcal{N}$ and $\mathcal{D}$ are compact. Thus, the hypotheses of Lemma 2.4 of Newey and McFadden [27] are satisfied. Hence, by UWL, the first term on the RHS in $(A.2)$ is $o_p(1)$. The second term on the RHS in $(A.2)$ is $o_p(1)$ by continuity of the expectation in $\eta \in \mathcal{N}$ and $\delta \in \mathcal{D}$, $\hat{\eta} \xrightarrow{p} \eta_0$ and $\dot{\delta} \xrightarrow{p} \delta_1$.

Then, by MAR,

$$\Big|\Big|\frac{1}{n}\sum_{i=1}^{n}\frac{d_iG_1(\hat{t}_i'\dot{\delta})}{G(\hat{t}_i'\dot{\delta})^2}\hat{t}_i\hat{t}_i' - \mathbb{E}\Big[\frac{G_1}{G}tt'\Big]\Big|\Big| = o_p(1).$$

Thus,

$$\hat{\delta} - \delta_1 = \mathbb{E}\Big[\frac{G_1}{G}tt'\Big]^{-1}\frac{1}{n}\sum_{i=1}^{n}\Big(\frac{d_i}{G(\hat{t}_i'\delta_1)} - 1\Big)\hat{t}_i + o_p(n^{-\frac{1}{2}}). \quad (A.3)$$

By a Taylor expansion around η_0, for some $\dot{\eta}$ on the line segment joining $\hat{\eta}$ and η_0,

$$\frac{1}{n}\sum_{i=1}^{n}\Big(\frac{d_i}{G(\hat{t}_i'\delta_1)} - 1\Big)\hat{t}_i = \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{d_i}{G_i} - 1\Big)t_i - \Big(\frac{1}{n}\sum_{i=1}^{n}\frac{d_iG_1(t_i(\dot{\eta})'\delta_1)}{G(t_i(\dot{\eta})'\delta_1)^2}t_i(\dot{\eta})\delta_1'\frac{\partial t_i(\dot{\eta})}{\partial \eta}\Big)(\hat{\eta} - \eta_0) + \Big(\frac{1}{n}\sum_{i=1}^{n}\Big(\frac{d_i}{G(t_i(\dot{\eta})'\delta_1)} - 1\Big)\frac{\partial t_i(\dot{\eta})}{\partial \eta}\Big)(\hat{\eta} - \eta_0) \quad (A.4)$$

By Assumptions 3.1, 3.2 and CS, for any $\eta \in \mathcal{N}$, $||dG_1(t(\eta)'\delta_1)t(\eta)\delta_1'(\partial t(\eta)/\partial \eta)/G(t(\eta)'\delta_1)^2|| \leq ||\delta_1||\kappa_1 b_t(x)b_{\partial t}(x)/\kappa^2$, where the last term on the RHS is bounded in expectation. Also, $G(t(\eta)'\delta_1)$, $G_1(t(\eta)'\delta_1)$, $t(\eta)$ and $\partial t(\eta)/\partial \eta$ are continuous in $\eta \in \mathcal{N}$, and $\mathcal{N}$ is compact. Thus, the hypotheses of Lemma 2.4 of Newey and McFadden [27] are satisfied. Hence, by UWL

$$\Big|\Big|\frac{1}{n}\sum_{i=1}^{n}\frac{d_iG_1(t_i(\dot{\eta})'\delta_1)}{G(t_i(\dot{\eta})'\delta_1)^2}t_i(\dot{\eta})\delta_1'\frac{\partial t_i(\dot{\eta})}{\partial \eta} - \mathbb{E}\Big[\frac{dG_1(t(\dot{\eta})'\delta_1)}{G(t(\dot{\eta})'\delta_1)^2}t(\dot{\eta})\delta_1'\frac{\partial t(\dot{\eta})}{\partial \eta}\Big]\Big|\Big| = o_p(1).$$

Then, by continuity of the expectation in $\eta \in \mathcal{N}$, $\hat{\eta} \xrightarrow{p} \eta_0$, T and MAR,

$$\Big|\Big|\frac{1}{n}\sum_{i=1}^{n}\frac{d_iG_1(t_i(\dot{\eta})'\delta_1)}{G(t_i(\dot{\eta})'\delta_1)^2}t_i(\dot{\eta})\delta_1'\frac{\partial t_i(\dot{\eta})}{\partial \eta} - \mathbb{E}\Big[\frac{G_1}{G}t\delta_1'\frac{\partial t}{\partial \eta}\Big]\Big|\Big| = o_p(1).$$

Similarly,

$$\Big|\Big|\frac{1}{n}\sum_{i=1}^{n}\Big(\frac{d_i}{G(t_i(\dot{\eta})'\delta_1)}-1\Big)\frac{\partial t_i(\dot{\eta})}{\partial \eta}\Big|\Big| = o_p(1).$$

Thus, from $(A.4)$ and $\hat{\eta}-\eta_0 = O_p(n^{-\frac{1}{2}})$,

$$\frac{1}{n}\sum_{i=1}^{n}\Big(\frac{d_i}{G(\hat{t}_i'\delta_1)}-1\Big)\hat{t}_i = \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{d_i}{G_i}-1\Big)t_i - \mathbb{E}\Big[\frac{G_1}{G}t\delta_1'\frac{\partial t}{\partial \eta}\Big](\hat{\eta}-\eta_0)+o_p(n^{-\frac{1}{2}}).$$

Substituting into $(A.3)$, using $(A.1)$,

$$\hat{\delta}-\delta_1 = \mathbb{E}\Big[\frac{G_1}{G}tt'\Big]^{-1}\Big(\frac{1}{n}\sum_{i=1}^{n}\Big(\frac{d_i}{G_i}-1\Big)t_i+\mathbb{E}\Big[\frac{G_1}{G}t\delta_1'\frac{\partial t}{\partial \eta}\Big]G_\eta^{-1}\frac{1}{n}\sum_{i=1}^{n}g_\eta(w_i,\eta_0)\Big)+o_p(n^{-\frac{1}{2}}). \quad (A.5)$$

The IPT estimating equation for γ_0 is

$$\frac{1}{n}\sum_{i=1}^{n}\frac{d_i\psi(z_i,\hat{\gamma})}{G(\hat{t}_i'\hat{\delta})}=0.$$

By a Taylor expansion around γ_0 and δ_1, for some $\dot{\gamma}$ on the line segment joining $\hat{\gamma}$ and γ_0, and for some $\ddot{\delta}$ on the line segment joining $\hat{\delta}$ and δ_1,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{d_i\psi_i}{G(\hat{t}_i'\delta_1)}-\frac{1}{n}\sum_{i=1}^{n}\frac{d_iG_1(\hat{t}_i'\ddot{\delta})}{G(\hat{t}_i'\ddot{\delta})^2}\psi_i\hat{t}_i'(\hat{\delta}-\delta_1)+\frac{1}{n}\sum_{i=1}^{n}\frac{d_i\Psi(z_i,\dot{\gamma})}{G(\hat{t}_i'\delta_1)}(\hat{\gamma}-\gamma_0)=o_p(n^{-\frac{1}{2}}). \quad (A.6)$$

By identical arguments to those used above, from consistency of $\hat{\eta}$ and $\hat{\delta}$,

$$\Big|\Big|\frac{1}{n}\sum_{i=1}^{n}\frac{d_iG_1(\hat{t}_i'\ddot{\delta})}{G(\hat{t}_i'\ddot{\delta})^2}\psi_i\hat{t}_i'-\mathbb{E}\Big[\frac{G_1}{G}\psi t'\Big]\Big|\Big|=o_p(1). \quad (A.7)$$

By T,

$$\begin{aligned}\Big|\Big|\frac{1}{n}\sum_{i=1}^{n}\frac{d_i\Psi_i(\dot{\gamma})}{G(\hat{t}_i'\delta_1)}-\mathbb{E}\Big[\frac{d\Psi}{G}\Big]\Big|\Big| \;\leq\; & \Big|\Big|\frac{1}{n}\sum_{i=1}^{n}\frac{d_i\Psi_i(\dot{\gamma})}{G(\hat{t}_i'\delta_1)}-\mathbb{E}\Big[\frac{d\Psi(\dot{\gamma})}{G(t(\hat{\eta})'\delta_1)}\Big]\Big|\Big| \\ & +\Big|\Big|\mathbb{E}\Big[\frac{d\Psi(\dot{\gamma})}{G(t(\hat{\eta})'\delta_1)}\Big]-\mathbb{E}\Big[\frac{d\Psi}{G}\Big]\Big|\Big|. \quad (A.8)\end{aligned}$$

By Assumptions 2.1, 3.1, 3.2 and CS, for any $\eta\in\mathcal{N}$ and $\gamma\in\Gamma$, $||d\Psi(\gamma)/G(t(\eta)'\delta_1)||\leq b_\Psi(z)/\kappa$, where the term on the RHS is bounded in expectation. Also, $\Psi(\gamma)$ and $G(t(\eta)'\delta_1)$ are continuous in parameters $\eta\in\mathcal{N}$ and $\gamma\in\Gamma$, and $\mathcal{N}$ and Γ are compact. Thus, the hypotheses of Lemma 2.4 of Newey and McFadden [27] are satisfied, and the first term on the RHS in $(A.8)$ is $o_p(1)$ by UWL. By continuity of the expectation in $\eta\in\mathcal{N}$ and $\gamma\in\Gamma$, $\hat{\eta}\xrightarrow{p}\eta_0$

and $\hat{\gamma} \xrightarrow{p} \gamma_0$, the second term on the RHS in $(A.8)$ is $o_p(1)$. Thus, by MAR and T,

$$\Big|\Big|\frac{1}{n}\sum_{i=1}^{n}\frac{d_i\Psi_i(\dot{\gamma})}{G(\hat{t}_i'\delta_1)} - \Psi\Big|\Big| = o_p(1). \tag{A.9}$$

Substituting $(A.7)$ and $(A.9)$ into $(A.6)$,

$$\Psi(\hat{\gamma}-\gamma_0) = -\frac{1}{n}\sum_{i=1}^{n}\frac{d_i\psi_i}{G(\hat{t}_i'\delta_1)} + \mathbb{E}\Big[\frac{G_1}{G}\psi t'\Big](\hat{\delta}-\delta_1) + o_p(n^{-\frac{1}{2}}). \tag{A.10}$$

By a Taylor expansion around η_0, for some $\ddot{\eta}$ on the line segment joining $\hat{\eta}$ and η_0,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{d_i\psi_i}{G(\hat{t}_i'\delta_1)} = \frac{1}{n}\sum_{i=1}^{n}\frac{d_i\psi_i}{G_i} - \frac{1}{n}\sum_{i=1}^{n}\frac{d_iG_1(t_i(\ddot{\eta})'\delta_1)}{G(t_i(\ddot{\eta})'\delta_1)^2}\psi_i\delta_1'\frac{\partial t_i(\ddot{\eta})}{\partial\eta}(\hat{\eta}-\eta_0).$$

By identical arguments to those used above,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{d_iG_1(t_i(\ddot{\eta})'\delta_1)}{G(t_i(\ddot{\eta})'\delta_1)^2}\psi_i\delta_1'\frac{\partial t_i(\ddot{\eta})}{\partial\eta} \xrightarrow{p} \mathbb{E}\Big[\frac{G_1}{G}\psi\delta_1'\frac{\partial t}{\partial\eta}\Big].$$

Thus, using $(A.1)$,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{d_i\psi_i}{G(\hat{t}_i'\delta_1)} = \frac{1}{n}\sum_{i=1}^{n}\frac{d_i\psi_i}{G_i} + \mathbb{E}\Big[\frac{G_1}{G}\psi\delta_1'\frac{\partial t}{\partial\eta}\Big]G_\eta^{-1}\frac{1}{n}\sum_{i=1}^{n}g_{\eta i} + o_p(n^{-\frac{1}{2}}). \tag{A.11}$$

Using $(A.5)$, $(A.10)$ and $(A.11)$, $\hat{\gamma}$ satisfies the expansion

$$\begin{aligned}
\Psi(\hat{\gamma}-\gamma_0) &= -\frac{1}{n}\sum_{i=1}^{n}\frac{d_i\psi_i}{G_i} + \mathbb{E}\Big[\frac{G_1}{G}\psi t'\Big]\mathbb{E}\Big[\frac{G_1}{G}tt'\Big]^{-1}\frac{1}{n}\sum_{i=1}^{n}\Big(\frac{d_i}{G_i}-1\Big)t_i \\
&\quad -\Big(\mathbb{E}\Big[\frac{G_1}{G}\psi\delta_1'\frac{\partial t}{\partial\eta}\Big] - \mathbb{E}\Big[\frac{G_1}{G}\psi t'\Big]\mathbb{E}\Big[\frac{G_1}{G}tt'\Big]^{-1}\mathbb{E}\Big[\frac{G_1}{G}t\delta_1'\frac{\partial t}{\partial\eta}\Big]\Big)G_\eta^{-1}\frac{1}{n}\sum_{i=1}^{n}g_{\eta i} \\
&\quad +o_p(n^{-\frac{1}{2}}) \\
&= -[I,\, -EF^{-1},\, CG_\eta^{-1}]\frac{1}{n}\sum_{i=}^{n}\begin{bmatrix}\frac{d_i\psi_i}{G_i}\\ \Big(\frac{d_i}{G_i}-1\Big)t_i \\ g_{\eta i}\end{bmatrix} + o_p(n^{-\frac{1}{2}}).
\end{aligned}$$

When $\mathbb{E}[\psi(z,\gamma_0)|x] \neq \Pi_0^\star t^\star(x,\eta_0)$,

$$\sqrt{n}(\hat{\gamma}-\gamma_0) \xrightarrow{d} \mathcal{N}\big(0, \Psi^{-1}\Sigma_1\Psi^{-1\prime}\big)$$

where $\Sigma_1 = \mathcal{V}_{(1)} - \mathcal{V}_{(2)} + \mathcal{V}_{(3)}$ for $\mathcal{V}_{(1)} = \mathbb{E}[\psi\psi'/G]$, $\mathcal{V}_{(2)} = \mathbb{E}[\psi t']\mathbb{E}[(G/(1-G))tt']^{-1}\mathbb{E}[\psi t']'$ and $\mathcal{V}_{(3)} = CG_\eta^{-1}B_\eta' + B_\eta G_\eta^{-1}C' + CG_\eta^{-1}\Omega_\eta G_\eta^{-1}C'$ where $B_\eta = \mathbb{E}[\psi g_\eta'] - EF^{-1}\mathbb{E}[(dG^{-1}-1)tg_\eta'].\square$

Part (b) - Variance structure under correct specification

When $\mathbb{E}[\psi(z,\gamma_0)|x] = \Pi_0^\star t^\star(x,\eta_0)$ where $\Pi_0 = (\Pi_0^{\star\prime}, 0')'$. By LIE,

$$\begin{aligned} C &= \mathbb{E}\Big[\frac{G_1}{G}\mathbb{E}[\psi|x]\delta_1'\frac{\partial t}{\partial \eta}\Big] - \mathbb{E}\Big[\frac{G_1}{G}\mathbb{E}[\psi|x]t'\Big]\mathbb{E}\Big[\frac{G_1}{G}tt'\Big]^{-1}\mathbb{E}\Big[\frac{G_1}{G}t\delta_1'\frac{\partial t}{\partial \eta}\Big] \\ &= \Pi_0\mathbb{E}\Big[\frac{G_1}{G}t\delta_1'\frac{\partial t}{\partial \eta}\Big] - \Pi_0\mathbb{E}\Big[\frac{G_1}{G}tt'\Big]\mathbb{E}\Big[\frac{G_1}{G}tt'\Big]^{-1}\mathbb{E}\Big[\frac{G_1}{G}t\delta_1'\frac{\partial t}{\partial \eta}\Big] \\ &= 0 \end{aligned}$$

Therefore,

$$\sqrt{n}(\hat{\gamma} - \gamma_0) \xrightarrow{d} \mathcal{N}\big(0, \Psi^{-1}\Sigma_0\Psi'^{-1}\big)$$

where $\Sigma_0 = \mathcal{V}_{(1)} - \mathcal{V}_{(2)}$ with $\mathcal{V}_{(1)} = \mathbb{E}[\psi\psi'/G]$ and $\mathcal{V}_{(2)} = \mathbb{E}[\psi t']\mathbb{E}[(G/(1-G))tt']^{-1}\mathbb{E}[\psi t']'$.

Now,

$$\begin{aligned} \Psi^{-1}(\mathcal{V}_{(1)} - \mathcal{V}_{(2)})\Psi'^{-1} &= \Psi^{-1}\left(\mathbb{E}\Big[\frac{\psi\psi'}{G}\Big] - \mathbb{E}[\psi t']\mathbb{E}\Big[\frac{G}{1-G}tt'\Big]^{-1}\mathbb{E}[\psi t']'\right)\Psi'^{-1} \\ &= \Psi^{-1}\left(\mathbb{E}\Big[\frac{\psi\psi'}{G}\Big] - \Pi_0^\star\mathbb{E}\Big[\frac{1-G}{G}t^\star t^{\star\prime}\Big]\Pi_0^{\star\prime}\right)\Psi'^{-1} \\ &= \Psi^{-1}\left(\mathbb{E}\Big[\frac{\psi\psi'}{G}\Big] - \mathbb{E}\Big[\frac{1-G}{G}q_0(x,\gamma_0)q_0(x,\gamma_0)'\Big]\right)\Psi'^{-1} \\ &= \Psi^{-1}\mathbb{E}\left[\frac{\mathbb{E}[\psi\psi'|x] - q_0(x,\gamma_0)q_0(x,\gamma_0)'}{G} + q_0(x,\gamma_0)q_0(x,\gamma_0)'\right]\Psi'^{-1} \end{aligned}$$

where the second line follows from Assumption 3.2 and $\Pi_0 t(x,\eta_0) = \Pi_0^\star t^\star(x,\eta_0)$, the third from writing $q_0(x,\gamma_0) = \mathbb{E}[\psi(z,\gamma_0)|x] = \Pi_0^\star t^\star(x,\eta_0)$ and the fourth by MAR. This is the asymptotic variance that corresponds the semiparametric efficiency lower bound in Section 3.2.1, i.e., the variance lower bound for an unconditional moment conditions model with MAR, as in Graham [11], based on Assumptions 2.1-2.5 and 3.1. □

Part (d) - Double robustness

Condition A.1 (Consistency under $p_0(X) = G(r(X)'\delta_0)$ and $\mathbb{E}[\psi(z,\gamma_0)|x] \neq \Pi_0^\star t^\star(x,\eta_0)$). The proof in Appendix B holds with $\hat{\pi}_i = \frac{1}{n}$ for all $i = 1,...,n$; note that the property $\mathbb{E}[\psi(z,\gamma_0)|x] = \Pi_0^\star t^\star(x,\eta_0)$ is not needed for the proof of consistency.

Condition A.2 (Consistency under $\mathbb{E}[\psi(z,\gamma_0)|x] = \Pi_0^\star t^\star(x,\eta_0)$ and if there is no δ_0 such that $p_0(X) = G(r(X)'\delta_0)$).
Suppose $\delta_\star$ is the unique solution to

$$\mathbb{E}\Big[\Big(\frac{d}{G(t'\delta)} - 1\Big)t\Big] = 0$$

Then, by similar arguments as used initially for the consistency of $\hat{\delta}$ in Appendix B, with $\delta_\star$ the minimiser of $Q_0(\delta) = ||\mathbb{E}[(d/G(t'\delta) - 1)t]||$, $\hat{\delta} \xrightarrow{p} \delta_\star$. Hence $\hat{\gamma}$ is a consistent estimator for

the $\gamma \in \Gamma$ that solves

$$\mathbb{E}\Big[\frac{d}{G(t'\delta_\star)}\psi(\gamma)\Big] = 0.$$

Now

$$\begin{aligned}
\mathbb{E}\Big[\frac{d}{G(t'\delta_\star)}\psi(\gamma)\Big] &= \mathbb{E}\Big[\frac{d}{G(t'\delta_\star)}\psi(\gamma)\Big] - \mathbb{E}[\psi(\gamma_0)] \\
&= \mathbb{E}\Big[\frac{d}{G(t'\delta_\star)}\psi(\gamma)\Big] - \Pi_0\mathbb{E}[t] \\
&= \mathbb{E}\Big[\frac{d}{G(t'\delta_\star)}\psi(\gamma)\Big] - \Pi_0\mathbb{E}\Big[\frac{d}{G(t'\delta_\star)}t\Big] \\
&= \mathbb{E}\Big[\frac{d}{G(t'\delta_\star)}(\psi(\gamma) - \psi(\gamma_0))\Big]
\end{aligned}$$

where the first equality follows from $\mathbb{E}[\psi(z, \gamma_0)] = 0$, the second by LIE and the third from the moment balancing equation used to estimate $\delta_\star$

$$\mathbb{E}\Big[\frac{d}{G(t'\delta_\star)}t\Big] = \mathbb{E}[t].$$

Therefore, $\hat{\gamma} \xrightarrow{p} \gamma_0$ by continuity of the moment function in $\gamma \in \Gamma$ and

$$\mathbb{E}\Big[\frac{d}{G(t'\delta_\star)}(\psi(\gamma) - \psi(\gamma_0))\Big] = 0.$$

□

Proof of Proposition 3.2

Part (a) - Consistency

The proof for consistency is given in Appendix B; the arguments from the proof hold throughout when $\hat{\pi}_i = 1/n$, $(i = 1, ..., n)$ is assumed. □

Part (c) - Variance structure under misspecification

Let $w = (d, z)$. Suppose functions $t(x, \eta_0)$ contain $\eta_0 := \eta_0(x_2) = \mathbb{E}[z_2|x_2]$. Suppose $\hat{\eta} := \hat{\eta}(x_2)$ is a nonparametric estimator for $\mathbb{E}[z_2|x_2]$. The estimating equation for δ_1 is

$$\frac{1}{n}\sum_{i=1}^{n}\Big(\frac{d_i}{G(\hat{t}_i'\hat{\delta})} - 1\Big)\hat{t}_i = 0.$$

By the same arguments used to derive equation $(A.3)$,

$$\hat{\delta} - \delta_1 = \mathbb{E}\Big[\frac{G_1}{G}tt'\Big]^{-1}\frac{1}{n}\sum_{i=1}^{n}\Big(\frac{d_i}{G(\hat{t}_i'\delta_1)} - 1\Big)\hat{t}_i + o_p(n^{-\frac{1}{2}}).$$

Nonparametric nuisance estimation of $\eta_0(x_2) = \mathbb{E}[z_2|x_2]$ must be taken into account in asymptotic variance calculations for $\hat{\delta}$. Since $\eta_0(x_2)$ is a conditional expectation, by Proposition 4

of Newey ([25], p.1361), the adjustment term $a_2(w_i)$ which embodies the effect of estimating $\eta_0(x_2)$ on $\sum_{i=1}^n((d_i/G(\hat{t}_i'\delta_1))-1)\hat{t}_i/n$ satisfies

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(\frac{d_i}{G(\hat{t}_i'\delta_1)}-1\Big)\hat{t}_i=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big\{\Big(\frac{d_i}{G_i}-1\Big)t_i+a_2(w_i)\Big\}+o_p(1),$$

where $a_2(w)=-d_2(x_2)(z_2-n_0(x_2))$ with

$$\begin{aligned}
d_2(x_2) &= \mathbb{E}\Big[\frac{dG_1}{G^2}t\delta_1'\frac{\partial t}{\partial\eta}-\Big(\frac{d}{G}-1\Big)t\Big|x_2\Big] \\
&= \mathbb{E}\Big[\frac{\mathbb{E}[d|y,x]G_1}{G^2}t\delta_1'\frac{\partial t}{\partial\eta}\Big|x_2\Big]-\mathbb{E}\Big[\Big(\frac{\mathbb{E}[d|y,x]}{G}-1\Big)t\Big|x_2\Big] \\
&= \mathbb{E}\Big[\frac{G_1}{G}t\delta_1'\frac{\partial t}{\partial\eta}\Big|x_2\Big]
\end{aligned}$$

by LIE and MAR.

Therefore,

$$\hat{\delta}-\delta_1 = \mathbb{E}\Big[\frac{G_1}{G}tt'\Big]^{-1}\Big\{\frac{1}{n}\sum_{i=1}^{n}\Big(\frac{d_i}{G_i}-1\Big)t_i-d_2(x_{2i})(z_{2i}-n_0(x_{2i}))\Big\}+o_p(n^{-\frac{1}{2}}). \quad \text{(A.12)}$$

Using the same arguments used to derive $(A.10)$,

$$\Psi(\hat{\gamma}-\gamma_0)=-\frac{1}{n}\sum_{i=1}^{n}\frac{d_i\psi_i}{G(\hat{t}_i'\delta_1)}+\mathbb{E}\Big[\frac{G_1}{G}\psi t'\Big](\hat{\delta}-\delta_1)+o_p(n^{-\frac{1}{2}}).$$

Similarly, by Proposition 4 of Newey ([25], p.1361), and using $(A.12)$,

$$\begin{aligned}
\Psi(\hat{\gamma}-\gamma_0) &= -\frac{1}{n}\sum_{i=1}^{n}\frac{d_i\psi_i}{G_i}+\mathbb{E}\Big[\frac{G_1}{G}\psi t'\Big]\Big[\frac{G_1}{G}tt'\Big]^{-1}\frac{1}{n}\sum_{i=1}^{n}\Big(\frac{d_i}{G_i}-1\Big)t_i \\
&\quad +\frac{1}{n}\sum_{i=1}^{n}\Big(d_1(x_{2i})-\mathbb{E}\Big[\frac{G_1}{G}\psi t'\Big]\Big[\frac{G_1}{G}tt'\Big]^{-1}d_2(x_{2i})\Big)(z_{2i}-\eta_o(x_{2i}))+o_p(n^{-\frac{1}{2}}),
\end{aligned}$$

where $d_1(x_2)=\mathbb{E}\Big[\frac{G_1}{G}\psi\delta_1'\frac{\partial t}{\partial\eta}\Big|x_2\Big]$ is an adjustment term that accounts for nonparametric conditional expectation estimation.

Let $v(x_2)=[d_1(x_2)-EF^{-1}d_2(x_2)]$. Then, by CLT, $\Psi\sqrt{n}(\hat{\gamma}-\gamma_0)\xrightarrow{d}\mathcal{N}(0,\Sigma_{NP})$ where

$$\begin{aligned}
\Sigma_{NP} &= \begin{pmatrix} I & -EF^{-1} & -I \end{pmatrix}\times \\
&\quad \begin{pmatrix} \mathbb{E}\Big[\frac{\psi\psi'}{G}\Big] & \mathbb{E}\Big[\frac{1-G}{G}\psi t'\Big] & \mathbb{E}[\psi v(x_2)'(z_2-\eta_0(x_2))] \\ \mathbb{E}\Big[\frac{1-G}{G}t\psi'\Big] & \mathbb{E}\Big[\frac{1-G}{G}tt'\Big] & \mathbb{E}\Big[tv(x_2)'\Big(\frac{d}{G}-1\Big)(z_2-\eta_0(x_2))\Big] \\ \mathbb{E}[\psi v(x_2)'(z_2-\eta_0(x_2))]' & \mathbb{E}\Big[tv(x_2)'\Big(\frac{d}{G}-1\Big)(z_2-\eta_0(x_2))\Big]' & \mathbb{E}[v(x_2)v(x_2)'(z_2-\eta_0(x_2))^2] \end{pmatrix}\begin{pmatrix} I \\ F^{-1}E' \\ -I \end{pmatrix}.
\end{aligned}$$

This leads to the following variance structure of Σ_{NP}.

$$\Sigma_{NP} = \mathcal{V}_{(1)} - \mathcal{V}_{(2)} + \mathcal{V}_{(NP)}$$

where

$$\begin{aligned}
\mathcal{V}_{(NP)} &= \mathbb{E}[v(x_2)v(x_2)'(z_2 - \eta_0(x_2))^2] \\
&\quad -\Big\{\mathbb{E}[\psi v(x_2)'(z_2 - \eta_0(x_2))] - EF^{-1}\mathbb{E}\Big[tv(x_2)'\Big(\frac{d}{G} - 1\Big)(z_2 - \eta_0(x_2))\Big]\Big\} \\
&\quad -\Big\{\mathbb{E}[\psi v(x_2)'(z_2 - \eta_0(x_2))] - EF^{-1}\mathbb{E}\Big[tv(x_2)'\Big(\frac{d}{G} - 1\Big)(z_2 - \eta_0(x_2))\Big]\Big\}'.
\end{aligned}$$

□

Part (b) - Efficiency under correct specification and when $x_2 = x$

Note when $x_2 = x$, the above variance structure holds with

$$\begin{aligned}
v(x) &= \mathbb{E}\Big[\frac{G_1}{G}\psi\delta_1'\frac{\partial t}{\partial \eta}\Big|x\Big] - \mathbb{E}\Big[\frac{G_1}{G}\psi t'\Big]\Big[\frac{G_1}{G}tt'\Big]^{-1}\mathbb{E}\Big[\frac{G_1}{G}t\delta_1'\frac{\partial t}{\partial \eta}\Big|x\Big] \\
&= \frac{G_1}{G}\mathbb{E}[\psi|x]\delta_1'\frac{\partial t}{\partial \eta} - \mathbb{E}\Big[\frac{G_1}{G}\mathbb{E}[\psi|x]t'\Big]\Big[\frac{G_1}{G}tt'\Big]^{-1}\frac{G_1}{G}t\delta_1'\frac{\partial t}{\partial \eta} \\
&= \Pi_0\frac{G_1}{G}t\delta_1'\frac{\partial t}{\partial \eta} - \Pi_0\mathbb{E}\Big[\frac{G_1}{G}tt'\Big]\mathbb{E}\Big[\frac{G_1}{G}tt'\Big]^{-1}\frac{G_1}{G}t\delta_1'\frac{\partial t}{\partial \eta} \\
&= 0,
\end{aligned}$$

where the second equality follows by LIE. Therefore, $\Sigma_{NP} = \mathcal{V}_{(1)} - \mathcal{V}_{(2)}$. This is equivalent to the variance lower bound of Robins et al. [34], also see Theorem 2.1 of Graham [11].

Part (d) - Double robustness

A double robustness result follows by identical arguments used for the Proof of Part (d) of Proposition 3.1. Note that only the consistency of $\hat{\eta}(x_2)$ for $\eta_0(x_2)$ is needed for the derivations, which follows under suitable regularity conditions for nonparametric estimation of $\eta_0(x_2)$. □

Proof of Proposition 3.3

Part (a) - Consistency

The proof for consistency is given in Appendix B. □

Part (c) - Variance structure under misspecification

By Lemma A1 of Ramalho and Smith [32],

$$\hat{\pi}_i = n^{-1} + n^{-1}\hat{g}_{\eta i}'\hat{\lambda}(1 + o_p(1)) + O_p(n^{-\frac{3}{2}})'\hat{\lambda} \tag{A.13}$$

uniformly in i, $(i = 1, ..., n)$. The GEL-weighted IPT estimating equation for γ_0 is

$$\sum_{i=1}^{n} \hat{\pi}_i \frac{d_i}{\hat{G}_i} \psi_i(\hat{\gamma}) = 0.$$

A Taylor expansion around $\hat{\gamma} = \gamma_0$ yields

$$\sum_{i=1}^{n} \hat{\pi}_i \frac{d_i}{\hat{G}_i} \psi_i + \sum_{i=1}^{n} \hat{\pi}_i \frac{d_i}{\hat{G}_i} \Psi_i(\dot{\gamma})(\hat{\gamma} - \gamma_0) = 0$$

for some $\dot{\gamma}$ on the line segment joining $\hat{\gamma}$ and γ_0. By identical arguments used in the Proof for consistency in Appendix B,

$$\Big|\Big| \sum_{i=1}^{n} \hat{\pi}_i \frac{d_i}{\hat{G}_i} \Psi_i(\dot{\gamma}) - \frac{1}{n} \sum_{i=1}^{n} \frac{d_i}{\hat{G}_i} \Psi_i(\dot{\gamma}) \Big|\Big| \leq o_p(1).$$

By similar arguments used to establish $(A.9)$,

$$\frac{1}{n} \sum_{i=1}^{n} \frac{d_i}{\hat{G}_i} \Psi_i(\dot{\gamma}) \xrightarrow{p} \Psi.$$

Therefore,

$$(\Psi + o_p(1))(\hat{\gamma} - \gamma_0) = -\sum_{i=1}^{n} \hat{\pi}_i \frac{d_i}{\hat{G}_i} \psi_i. \tag{A.14}$$

For any $\eta \in \mathcal{N}$ and $\delta \in \mathcal{D}$, by CS, $||d\psi(z, \gamma_0)/G(t(\eta)'\delta)|| \leq b_\psi(z)/\kappa$. Since $\mathbb{E}[b_\psi(z)] < \infty$ by Assumption 2.1, $||n^{-1} \sum_{i=1}^{n} d_i \psi_i / G(t_i(\eta)'\delta)|| = O_p(1)$. Hence, by CS, noting $||\hat{\lambda}|| = O_p(n^{-\frac{1}{2}})$ (Cf. Theorem 3.1 of Newey and Smith [28], p.226), from $(A.14)$ and using $(A.13)$,

$$\begin{aligned} \sum_{i=1}^{n} \hat{\pi}_i \frac{d_i}{\hat{G}_i} \psi_i &= \frac{1}{n} \sum_{i=1}^{n} \frac{d_i}{\hat{G}_i} \psi_i + \frac{1}{n} \sum_{i=1}^{n} \frac{d_i}{\hat{G}_i} \psi_i \hat{g}'_{\eta i} \hat{\lambda}(1 + o_p(1)) + \frac{1}{n} \sum_{i=1}^{n} \frac{d_i}{\hat{G}_i} \psi_i O_p(n^{-\frac{1}{2}})' \hat{\lambda} \\ &= \frac{1}{n} \sum_{i=1}^{n} \frac{d_i}{\hat{G}_i} \psi_i + \frac{1}{n} \sum_{i=1}^{n} \frac{d_i}{\hat{G}_i} \psi_i \hat{g}'_{\eta i} \hat{\lambda}(1 + o_p(1)) + O_p(n^{-1}). \end{aligned} \tag{A.15}$$

By T,

$$\begin{aligned} \Big|\Big| \frac{1}{n} \sum_{i=1}^{n} \frac{d_i}{\hat{G}_i} \psi_i \hat{g}'_{\eta i} - \mathbb{E}\Big[\frac{d}{G} \psi g'_\eta\Big] \Big|\Big| \leq & \Big|\Big| \frac{1}{n} \sum_{i=1}^{n} \frac{d_i}{\hat{G}_i} \psi_i \hat{g}'_{\eta i} - \mathbb{E}\Big[\frac{d}{G(t(\hat{\eta})'\hat{\delta})} \psi(z, \gamma_0) g_\eta(w, \hat{\eta})' \Big] \Big|\Big| \\ & + \Big|\Big| \mathbb{E}\Big[\frac{d}{G(t(\hat{\eta})'\hat{\delta})} \psi(z, \gamma_0) g_\eta(w, \hat{\eta})' \Big] - \mathbb{E}\Big[\frac{d}{G} \psi g'_\eta\Big] \Big|\Big| \end{aligned} \tag{A.16}$$

For any $\eta \in \mathcal{N}$ and $\delta \in \mathcal{D}$, by CS, $||d\psi(z, \gamma_0) g_\eta(w, \eta)'/G(t(\eta)'\delta)|| \leq b_\psi(z)||g_\eta(w, \eta)||/\kappa$, and it is assumed that $\mathbb{E}[b_\psi(z)||g_\eta(w, \eta)||] < \infty$. Moreover, $g_\eta(w, \eta)$ and $G(t(\eta)'\delta)$ are continuous in

parameters $\eta \in \mathcal{N}$ and $\delta \in \mathcal{D}$, and $\mathcal{N}$ and $\mathcal{D}$ are compact. Hence, the hypotheses of Lemma 2.4 of Newey and McFadden are satisfied. Thus, by UWL, the first term on the RHS in $(A.16)$ is $o_p(1)$. The second term on the RHS in $(A.16)$ is $o_p(1)$ by continuity of the expectation in $\eta \in \mathcal{N}$ and $\delta \in \mathcal{D}$, $\hat{\eta} \xrightarrow{p} \eta_0$ and $\hat{\delta} \xrightarrow{p} \delta_1$.

Then, by MAR,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{d_i}{\hat{G}_i}\psi_i\hat{g}'_{\eta i} \xrightarrow{p} \mathbb{E}[\psi g'_\eta]. \tag{A.17}$$

By a Taylor expansion around η_0 and δ_1, for some $\dot{\eta}$ on the line segment joining $\hat{\eta}$ and η_0, and for some $\dot{\delta}$ on the line segment joining $\hat{\delta}$ and δ_0,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\frac{d_i}{\hat{G}_i}\psi_i &= \frac{1}{n}\sum_{i=1}^{n}\frac{d_i}{G_i}\psi_i - \Big(\frac{1}{n}\sum_{i=1}^{n}\frac{d_i G_1(t_i'\dot{\delta})}{G(t_i'\dot{\delta})^2}\psi_i t_i'\Big)(\hat{\delta}-\delta_1)\\ &\quad -\Big(\frac{1}{n}\sum_{i=1}^{n}\frac{d_i G_1(t_i(\dot{\eta})'\delta_1)}{G(t_i(\dot{\eta})'\delta_1)^2}\psi_i\delta_1'\frac{\partial t_i(\dot{\eta})}{\partial\eta}\Big)(\hat{\eta}-\eta_0).\end{aligned} \tag{A.18}$$

By MAR and identical arguments used to establish $(A.7)$ and $(A.11)$,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\frac{d_i G_1(t_i(\dot{\eta})'\delta_1)}{G(t_i(\dot{\eta})'\delta_1)^2}\psi_i\delta_1'\frac{\partial t_i(\dot{\eta})}{\partial\eta} &\xrightarrow{p} \mathbb{E}\Big[\frac{G_1}{G}\psi\delta_1'\frac{\partial t}{\partial\eta}\Big]\\ \frac{1}{n}\sum_{i=1}^{n}\frac{d_i G_1(t_i'\dot{\delta})}{G(t_i'\dot{\delta})^2}\psi_i t_i' &\xrightarrow{p} \mathbb{E}\Big[\frac{G_1}{G}\psi t'\Big],\end{aligned}$$

hence, by $(A.18)$,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{d_i}{\hat{G}_i}\psi_i = \frac{1}{n}\sum_{i=1}^{n}\frac{d_i}{G_i}\psi_i - \mathbb{E}\Big[\frac{G_1}{G}\psi t'\Big](\hat{\delta}-\delta_1) - \mathbb{E}\Big[\frac{G_1}{G}\psi\delta_1'\frac{\partial t}{\partial\eta}\Big](\hat{\eta}-\eta_0). \tag{A.19}$$

By substitution of $(A.15)$, $(A.17)$ and $(A.19)$ into $(A.14)$,

$$(\Psi + o_p(1))(\hat{\gamma}-\gamma_0) = -\frac{1}{n}\sum_{i=1}^{n}\frac{d_i}{G_i}\psi_i - \mathbb{E}[\psi g'_\eta]\hat{\lambda} + \mathbb{E}\Big[\frac{G_1}{G}\psi t'\Big](\hat{\delta}-\delta_1) + \mathbb{E}\Big[\frac{G_1}{G}\psi\delta_1'\frac{\partial t}{\partial\eta}\Big](\hat{\eta}-\eta_0). \tag{A.20}$$

The GEL-weighted IPT estimating equation for δ_1 is

$$\sum_{i=1}^{n}\hat{\pi}_i\Big(\frac{d_i}{\hat{G}_i}-1\Big)\hat{t}_i = 0.$$

By a Taylor expansion around δ_1, for some $\dot{\delta}$ on the line segment joining $\hat{\delta}$ and δ_1,

$$\sum_{i=1}^{n}\hat{\pi}_i\Big(\frac{d_i}{G(\hat{t}_i'\delta_1)}-1\Big)\hat{t}_i - \Big(\sum_{i=1}^{n}\hat{\pi}_i\frac{d_i G_1(\hat{t}_i'\dot{\delta})}{G(\hat{t}_i'\dot{\delta})^2}\hat{t}_i\hat{t}_i'\Big)(\hat{\delta}-\delta_1) = 0. \tag{A.21}$$

For any $\eta \in \mathcal{N}$, $\delta \in \mathcal{D}$, $||(dG_1(t(\eta)'\delta))t(\eta)t(\eta)'/G(t(\eta)'\delta)^2|| \leq \kappa_1 b_t(x)^2/\kappa^2$, where $\mathbb{E}[b_t(x)^2] < \infty$ by Assumption 3.2. Hence by M,

$$\frac{1}{n}\sum_{i=1}^{n} \frac{d_i G_1(\hat{t}_i'\dot{\delta})}{G(\hat{t}_i'\dot{\delta})^2}\hat{t}_i\hat{t}_i' = O_p(1). \tag{A.22}$$

By Lemma A.1 of Newey and Smith, $\hat{\pi}_i = n^{-1} + o_p(1)$. Thus, by $(A.22)$ and CS,

$$\Big|\Big|\sum_{i=1}^{n} \hat{\pi}_i \frac{d_i G_1(\hat{t}_i'\dot{\delta})}{G(\hat{t}_i'\dot{\delta})^2}\hat{t}_i\hat{t}_i' - \frac{1}{n}\sum_{i=1}^{n} \frac{d_i G_1(\hat{t}_i'\dot{\delta})}{G(\hat{t}_i'\dot{\delta})^2}\hat{t}_i\hat{t}_i'\Big|\Big| = o_p(1). \tag{A.23}$$

By identical arguments used to establish $(A.3)$,

$$\frac{1}{n}\sum_{i=1}^{n} \frac{d_i G_1(\hat{t}_i'\dot{\delta})}{G(\hat{t}_i'\dot{\delta})^2}\hat{t}_i\hat{t}_i' \xrightarrow{p} \mathbb{E}\Big[\frac{G_1}{G}tt'\Big]. \tag{A.24}$$

Using $(A.13)$,

$$\begin{aligned}\sum_{i=1}^{n} \hat{\pi}_i\Big(\frac{d_i}{G(\hat{t}_i'\delta_1)} - 1\Big)\hat{t}_i &= \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{d_i}{G(\hat{t}_i'\delta_1)} - 1\Big)\hat{t}_i + \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{d_i}{G(\hat{t}_i'\delta_1)} - 1\Big)\hat{t}_i\hat{g}_{\eta i}'\hat{\lambda}(1 + o_p(1)) \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{d_i}{G(\hat{t}_i'\delta_1)} - 1\Big)\hat{t}_i O_p(n^{-\frac{1}{2}})'\hat{\lambda}. \end{aligned} \tag{A.25}$$

For any $\eta \in \mathcal{N}$, by CS, $||(dG(t(\eta)'\delta_1)^{-1} - 1)t(\eta)g_\eta(w,\eta)'|| \leq (\kappa^{-1}+1)b_t(x)||g_\eta(w,\eta)||$, and it is assumed that $\mathbb{E}[b_t(x)||g_\eta(w,\eta)||] < \infty$. Moreover, $G(t(\eta)'\delta_1)$, $t(\eta)$ and $g_\eta(w,\eta)$ are continuous in $\eta \in \mathcal{N}$, and $\mathcal{N}$ is compact. Thus the hypotheses of Lemma 2.4 of Newey and McFadden [27] are satisfied. Hence,

$$\Big|\Big|\frac{1}{n}\sum_{i=1}^{n}\Big(\frac{d_i}{G(\hat{t}_i'\delta_1)} - 1\Big)\hat{t}_i\hat{g}_{\eta i}' - \mathbb{E}\Big[\Big(\frac{d}{G(t(\hat{\eta})'\delta_1)} - 1\Big)t(\hat{\eta})g(w,\hat{\eta})'\Big]\Big|\Big| = o_p(1).$$

Moreover, by continuity of the expectation in $\eta \in \mathcal{N}$ and $\hat{\eta} \xrightarrow{p} \eta_0$,

$$\Big|\Big|\mathbb{E}\Big[\Big(\frac{d}{G(t(\hat{\eta})'\delta_1)} - 1\Big)t(\hat{\eta})g(w,\hat{\eta})'\Big] - \mathbb{E}\Big[\Big(\frac{d}{G} - 1\Big)tg_\eta'\Big]\Big|\Big| = o_p(1).$$

Also, for any $\eta \in \mathcal{N}$ and $\delta \in \mathcal{D}$, $||(dG(t(\eta)'\delta)^{-1} - 1)t(\eta)|| \leq (\kappa^{-1}+1)b_t(x)$, where $\mathbb{E}[b_t(x)] < \infty$ by Assumption 3.2. Thus, by M, CS, and noting $||\hat{\lambda}|| = O_p(n^{-\frac{1}{2}})$,

$$\Big|\Big|\frac{1}{n}\sum_{i=1}^{n}\Big(\frac{d_i}{G(\hat{t}_i'\delta_1)} - 1\Big)\hat{t}_i O_p(n^{-\frac{1}{2}})'\hat{\lambda}\Big|\Big| = O_p(n^{-1}).$$

Thus, by $(A.25)$,

$$\sum_{i=1}^{n}\hat{\pi}_i\Big(\frac{d_i}{G(\hat{t}_i'\delta_1)}-1\Big)\hat{t}_i \;=\; \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{d_i}{G(\hat{t}_i'\delta_1)}-1\Big)\hat{t}_i+\mathbb{E}\Big[\Big(\frac{d}{G}-1\Big)tg_\eta'\Big]\hat{\lambda}+O_p(n^{-1}).$$

Thus, by $(A.21)$, $(A.23)$, $(A.24)$ and $(A.25)$,

$$\Big(\mathbb{E}\Big[\frac{G_1}{G}tt'\Big]+o_p(1)\Big)(\hat{\delta}-\delta_1)=\frac{1}{n}\sum_{i=1}^{n}\Big(\frac{d_i}{G(\hat{t}_i'\delta_1)}-1\Big)\hat{t}_i+\mathbb{E}\Big[\Big(\frac{d}{G}-1\Big)tg_\eta'\Big]\hat{\lambda}+O_p(n^{-1}). \quad (A.26)$$

By a Taylor expansion around η_0, and by identical arguments used to establish $(A.5)$, since $\hat{\lambda}$ and $(\hat{\eta}-\eta_0)$ are $O_p(n^{-\frac{1}{2}})$ (cf. Theorem 3.2 of Newey and Smith [28]),

$$\begin{aligned}(\hat{\delta}-\delta_1) \;=\; & \mathbb{E}\Big[\frac{G_1}{G}tt'\Big]^{-1}\Big\{\frac{1}{n}\sum_{i=1}^{n}\Big(\frac{d_i}{G_i}-1\Big)t_i+\mathbb{E}\Big[\Big(\frac{d}{G}-1\Big)tg_\eta'\Big]\hat{\lambda}\\ & -\mathbb{E}\Big[\frac{G_1}{G}t\delta_1'\frac{\partial t}{\partial\eta}\Big](\hat{\eta}-\eta_0)\Big\}+o_p(n^{-\frac{1}{2}}). \quad (A.27)\end{aligned}$$

By substitution of $(A.27)$ into $(A.20)$,

$$\begin{aligned}(\Psi+o_p(1))(\hat{\gamma}-\gamma_0) \;=\; & -\frac{1}{n}\sum_{i=1}^{n}\frac{d_i}{G_i}\psi_i+\mathbb{E}\Big[\frac{G_1}{G}\psi t'\Big]\mathbb{E}\Big[\frac{G_1}{G}tt'\Big]^{-1}\frac{1}{n}\sum_{i=1}^{n}\Big(\frac{d_i}{G_i}-1\Big)t_i\\ & -\Big(\mathbb{E}[\psi g_\eta']-\mathbb{E}\Big[\frac{G_1}{G}\psi t'\Big]\mathbb{E}\Big[\frac{G_1}{G}tt'\Big]^{-1}\mathbb{E}\Big[\Big(\frac{d}{G}-1\Big)tg_\eta'\Big]\Big)\hat{\lambda}\\ & +\Big(\mathbb{E}\Big[\frac{G_1}{G}\psi\delta_1'\frac{\partial t}{\partial\eta}\Big]-\mathbb{E}\Big[\frac{G_1}{G}\psi t'\Big]\mathbb{E}\Big[\frac{G_1}{G}tt'\Big]^{-1}\mathbb{E}\Big[\frac{G_1}{G}t\delta_1'\frac{\partial t}{\partial\eta}\Big]\Big)(\hat{\eta}-\eta_0).\end{aligned}$$

Let $H_\eta=(G_\eta'\Omega_\eta^{-1}G_\eta)^{-1}G_\eta'\Omega_\eta^{-1}$, $P_\eta=\Omega_\eta^{-1}-\Omega_\eta^{-1}G_\eta'(G_\eta'\Omega_\eta^{-1}G_\eta)G_\eta^{-1}\Omega_\eta^{-1}$ for $\Omega_\eta=\mathbb{E}[g_\eta g_\eta']$, $B_\eta=\mathbb{E}[\psi g_\eta']-\mathbb{E}\Big[G_1\psi t'/G\Big]\mathbb{E}\Big[G_1tt'/G\Big]^{-1}\mathbb{E}\Big[(dG^{-1}-1)tg_\eta'\Big]$, and $C=\mathbb{E}\Big[\frac{G_1}{G}\psi\delta_1'\frac{\partial t}{\partial\eta}\Big]-\mathbb{E}\Big[\frac{G_1}{G}\psi t'\Big]$ $\times\mathbb{E}\Big[\frac{G_1}{G}tt'\Big]^{-1}\mathbb{E}\Big[\frac{G_1}{G}t\delta_1'\frac{\partial t}{\partial\eta}\Big]$. By Newey and Smith ([28], equation $(A.8)$, p.240), $\hat{\eta}-\eta_0=-H_\eta n^{-1}\sum_{i=1}^{n}g_{\eta i}+o_p(n^{-\frac{1}{2}})$ and $\hat{\lambda}=-P_\eta n^{-1}\sum_{i=1}^{n}g_{\eta i}+o_p(n^{-\frac{1}{2}})$. Let $E=\mathbb{E}\Big[G_1\psi t'/G\Big]$ and $F=\mathbb{E}\Big[G_1tt'/G\Big]$. Then,

$$\hat{\gamma}-\gamma_0 \;=\; -\Psi^{-1}(1,-EF^{-1},CH_\eta-B_\eta P_\eta)\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix}\frac{d_i}{G_i}\psi_i\\ \Big(\frac{d_i}{G_i}-1\Big)t_i\\ g_{\eta i}\end{pmatrix}+o_p(n^{-\frac{1}{2}}),$$

which leads to the following variance structure by CLT. Making use of the property that $H_\eta\Omega_\eta P_\eta=P_\eta\Omega_\eta H_\eta'=0$ and $P_\eta\Omega_\eta P_\eta=P_\eta$,

$$\sqrt{n}(\hat{\gamma}-\gamma_0)\xrightarrow{d}\mathcal{N}(0,\Psi^{-1}\Sigma_\star\Psi'^{-1})$$

where $\Sigma_\star = \Sigma_1 - B_\eta P_\eta B_\eta'$ where Σ_1 is the asymptotic variance for the IPT estimator from Proposition 3.1(c) with the slight modification, since here η_0 is overidentified, $\mathcal{V}_{(3)} = CH_\eta B_\eta + B_\eta H_\eta' C' + CH_\eta \Omega_\eta^{-1} H_\eta C'$. This would be the $\mathcal{V}_{(3)}$ term in Proposition 3.1(c) if two-step efficient GMM or GEL was used to estimate the overidentified system $\mathbb{E}[g_\eta(w, \eta_0)] = 0$. □

Part (b) - Variance structure under correct specification

As before in the Proof of Proposition 3.1(b), $C = 0$. Therefore $\Sigma_\star = \Sigma_0 - B_\eta P_\eta B_\eta'$. □

Part (d) - Double robustness

Condition A.1 (Consistency under $p_0(X) = G(r(X)'\delta_0)$ and $\mathbb{E}[\psi(z, \gamma_0)|x] \neq \Pi_0^\star t^\star(x, \eta_0)$). The proof of consistency in Appendix B holds since the condition $\mathbb{E}[\psi(z, \gamma_0)|x] = \Pi_0^\star t^\star(x, \eta_0)$ is not used in the proof.

Condition A.2 (Consistency under $\mathbb{E}[\psi(z, \gamma_0)|x] = \Pi_0^\star t^\star(x, \eta_0)$ and if there is no δ_0 such that $p_0(X) = G(r(X)'\delta_0)$).
For the reverse condition, it is shown in Appendix B that for any value of $\delta \in \mathcal{D}$ and $\gamma \in \Gamma$,

$$\begin{aligned} \Big|\Big| \sum_{i=1}^n \hat{\pi}_i \frac{D_i}{G(\hat{t}_i'\delta)} \psi_i(\gamma) - \mathbb{E}\Big[\frac{D}{G(t(\eta)'\delta)}\psi(z,\gamma)\Big]_{\eta=\hat{\eta}} \Big|\Big| &\leq o_p(1) \\ \Big|\Big| \sum_{i=1}^n \hat{\pi}_i \Big(\frac{D_i}{G(\hat{t}_i'\delta)} - 1\Big) - \mathbb{E}\Big[\Big(\frac{D}{G(t(\eta)'\delta)} - 1\Big)t(\eta)\Big]_{\eta=\hat{\eta}} \Big|\Big| &\leq o_p(1). \end{aligned}$$

The rest of the proof for the reverse condition is the same as for Condition A.2 of Proposition 3.1(c). □

B Consistency

The GEL implied probabilities are given by $\hat{\pi}_i = \hat{\rho}_{\eta i} / \sum_{j=1}^n \hat{\rho}_{\eta j}$ with $\hat{\rho}_{\eta i} = \rho(\hat{\lambda}'\hat{g}_{\eta i})$, $(i = 1, ..., n)$. First, for the consistency of $\hat{\delta}$, write

$$\sum_{i=1}^n \hat{\pi}_i \Big(\frac{d_i}{\hat{G}_i} - 1\Big) t_i(\hat{\eta}) = -\frac{1}{n} \sum_{i=1}^n \hat{\rho}_{\eta i} \Big(\frac{d_i}{\hat{G}_i} - 1\Big) t_i(\hat{\eta}) + R_{1n},$$

where

$$R_{1n} = \Big(\frac{n}{\sum_{j=1}^n \hat{\rho}_{\eta j}} + 1\Big) \frac{1}{n} \sum_{i=1}^n \hat{\rho}_{\eta i} \Big(\frac{d_i}{\hat{G}_i} - 1\Big) t_i(\hat{\eta}).$$

By Lemma A.1 of Newey and Smith [28], p.239, $\hat{\rho}_{\eta j} = -1 + o_p(1)$ uniformly $j = 1, ..., n$. Therefore,

$$\frac{n}{\sum_{j=1}^n \hat{\rho}_{\eta j}} + 1 = o_p(1).$$

By T, for any $\delta \in \mathcal{D}$,

$$\begin{aligned}
&\Big|\Big|\frac{1}{n}\sum_{i=1}^{n}\hat{\rho}_{\eta i}\Big(\frac{d_i}{G(t_i(\hat{\eta})'\delta)}-1\Big)t_i(\hat{\eta})-\frac{1}{n}\sum_{i=1}^{n}\Big(\frac{d_i}{G(t_i'\delta)}-1\Big)t_i\Big|\Big| \\
&\leq \Big|\Big|\frac{1}{n}\sum_{i=1}^{n}\hat{\rho}_{\eta i}\Big(\frac{d_i}{G(t_i(\hat{\eta})'\delta)}-1\Big)t_i(\hat{\eta})-\mathbb{E}\Big[\rho(\lambda' g_\eta(w,\eta))\Big(\frac{d}{G(t(\eta)'\delta)}-1\Big)t(\eta)\Big]_{(\lambda,\eta)=(\hat{\lambda},\hat{\eta})}\Big|\Big| \\
&+\Big|\Big|\mathbb{E}\Big[\rho(\lambda' g_\eta(w,\eta))\Big(\frac{d}{G(t(\eta)'\delta)}-1\Big)t(\eta)\Big]_{(\lambda,\eta)=(\hat{\lambda},\hat{\eta})}-\mathbb{E}\Big[\Big(\frac{d}{G(t'\delta)}-1\Big)t\Big]\Big|\Big| \\
&+\Big|\Big|\mathbb{E}\Big[\Big(\frac{d}{G(t'\delta)}-1\Big)t\Big]-\frac{1}{n}\sum_{i=1}^{n}\Big(\frac{d_i}{G(t_i'\delta)}-1\Big)t_i\Big|\Big|. \qquad (B.1)
\end{aligned}$$

By Lemma A.1 of Newey and Smith ([28], p.239), $\max_{1\leq i\leq n}\sup_{\eta\in\mathcal{N}}\sup_{\lambda\in\hat{\Lambda}_n(\eta)}|\lambda' g_i(\eta)| = o_p(1)$, so that $\hat{\rho}_{\eta i} = -1 + o_p(1)$ uniformly $i = 1, ..., n$. Thus, eventually, there is a constant $1 < C < \infty$ such that $|\hat{\rho}_{\eta i}| \leq C$, $(i = 1, ..., n)$. Thus, by CS and Assumption 3.2, for any $\eta \in \mathcal{N}$ and $\delta \in \mathcal{D}$, $||\rho(\lambda' g(w,\eta))(dG(t(\eta)'\delta)^{-1} - 1)t(\eta)|| \leq C(\kappa^{-1}+1)b_t(x)$ w.p.a.1, where $\mathbb{E}[b_t(x)] < \infty$. Moreover, by Assumption 1 of Newey and Smith ([28], p.226), $\rho(\lambda' g(w,\eta))$ is continuously differentiable in a neighbourhood of zero, $g(w,\eta)$ is continuous at each $\eta \in \mathcal{N}$ w.p.1, and $\mathcal{N}$ and $\hat{\Lambda}_n(\eta)$ are compact. Also, $t(\eta)$ is continuous in $\eta \in \mathcal{N}$, and $G(t(\eta)'\delta)$ is continuous in $\eta \in \mathcal{N}$ by Assumptions 3.1 and 3.2. Thus, the hypotheses of Lemma 2.4 of Newey and McFadden [27] are satisfied. Hence, by UWL, the first term on the RHS of $(B.1)$ is $o_p(1)$. Similarly, the third term on the RHS of $(B.1)$ is $o_p(1)$ by UWL for any $\delta \in \mathcal{D}$. The second term on the RHS of $(B.1)$ is $o_p(1)$ by continuity of the expectation in $\lambda \in \hat{\Lambda}_n(\eta)$ and $\eta \in \mathcal{N}$, $\hat{\lambda} \xrightarrow{p} 0$ and $\hat{\eta} \xrightarrow{p} \eta_0$, by Theorem 3.1 of Newey and Smith ([28], p.226). Therefore, the LHS of $(B.1)$ is $o_p(1)$.

Thus, by T, for any $\delta \in \mathcal{D}$,

$$\Big|\Big|\frac{1}{n}\sum_{i=1}^{n}\Big(\frac{d_i}{G(t_i'\delta)}-1\Big)t_i\Big|\Big| \leq \Big|\Big|\frac{1}{n}\sum_{i=1}^{n}\hat{\rho}_{\eta i}\Big(\frac{d_i}{G(t_i(\hat{\eta})'\delta)}-1\Big)t_i(\hat{\eta})\Big|\Big| + o_p(1).$$

From first order conditions for $\hat{\delta}$,

$$\Big|\Big|\frac{1}{n}\sum_{i=1}^{n}\hat{\rho}_{\eta i}\Big(\frac{d_i}{G(t_i(\hat{\eta})'\hat{\delta})}-1\Big)t_i(\hat{\eta})\Big|\Big| \leq o_p(1),$$

so that by T,

$$\Big|\Big|\frac{1}{n}\sum_{i=1}^{n}\Big(\frac{d_i}{G(t_i'\hat{\delta})}-1\Big)t_i\Big|\Big| \leq o_p(1). \qquad (B.2)$$

Now, let

$$Q_n(\delta) = \Big|\Big|\frac{1}{n}\sum_{i=1}^{n}\Big(\frac{d_i}{G(t_i'\delta)}-1\Big)t_i\Big|\Big| \text{ and } Q_0(\delta) = \Big|\Big|\mathbb{E}\Big[\Big(\frac{d}{G(t(x,\eta_0)'\delta)}-1\Big)t(x,\eta_0)\Big]\Big|\Big|$$

Similar arguments are now used to those of the Proof of Theorem 2.1 of Newey and McFadden ([27], p.2121-2), and Proof of Lemma A.2 of Otsu [29].

By uniqueness of δ_0 (and thus δ_1), for any $\epsilon > 0$, $\kappa = \inf_{\delta \in \mathcal{D}, ||\delta - \delta_1|| > \epsilon} Q_0(\delta) > 0$, and $Q_0(\delta_1) = 0$. Then, $||\hat{\delta} - \delta_1|| > \epsilon$ implies $Q_0(\hat{\delta}) \geq \inf_{\delta \in \mathcal{D}, |\delta - \delta_1| > \epsilon} Q_0(\delta) = \kappa > 0$. Hence,

$$\mathbb{P}(||\hat{\delta} - \delta_1|| > \epsilon) \leq \mathbb{P}(Q_0(\hat{\delta}) \geq \kappa). \tag{B.3}$$

Write

$$\begin{aligned}
\mathbb{P}(Q_0(\hat{\delta}) \geq \kappa) &= \mathbb{P}\Big(Q_0(\hat{\delta}) \geq \kappa \Big| \sup_{\delta \in \mathcal{D}} |Q_n(\delta) - Q_0(\delta)| \leq \frac{\kappa}{2}\Big)\mathbb{P}\Big(\sup_{\delta \in \mathcal{D}} |Q_n(\delta) - Q_0(\delta)| \leq \frac{\kappa}{2}\Big) \\
&\quad + \mathbb{P}\Big(Q_0(\hat{\delta}) \geq \kappa \Big| \sup_{\delta \in \mathcal{D}} |Q_n(\delta) - Q_0(\delta)| > \frac{\kappa}{2}\Big)\mathbb{P}\Big(\sup_{\delta \in \mathcal{D}} |Q_n(\delta) - Q_0(\delta)| > \frac{\kappa}{2}\Big) \\
&\leq \mathbb{P}\Big(\big\{Q_0(\hat{\delta}) \geq \kappa\big\} \cap \Big\{\sup_{\delta \in \mathcal{D}} |Q_n(\delta) - Q_0(\delta)| \leq \frac{\kappa}{2}\Big\}\Big) + \mathbb{P}\Big(\sup_{\delta \in \mathcal{D}} |Q_n(\delta) - Q_0(\delta)| > \frac{\kappa}{2}\Big).
\end{aligned}$$

For the first probability after the inequality, note that the event that $\big\{Q_0(\hat{\delta}) \geq \kappa\big\} \cap \Big\{ \sup_{\delta \in \mathcal{D}} |Q_n(\delta) - Q_0(\delta)| \leq \frac{\kappa}{2}\Big\}$ implies that the event $\big\{Q_n(\hat{\delta}) \geq \frac{\kappa}{2}\big\}$. Therefore,

$$\begin{aligned}
\mathbb{P}\Big(\big\{Q_0(\hat{\delta}) \geq \kappa\big\} \cap \Big\{\sup_{\delta \in \mathcal{D}} |Q_n(\delta) - Q_0(\delta)| \leq \frac{\kappa}{2}\Big\}\Big) &\leq \mathbb{P}\Big(Q_n(\hat{\delta}) \geq \frac{\kappa}{2}\Big) \\
&= o(1),
\end{aligned}$$

since $Q_n(\hat{\delta}) = o_p(1)$ by $(B.2)$.

In order to bound the third term on the RHS of $(B.1)$, it was shown by UWL that $\sup_{\delta \in \mathcal{D}} |Q_n(\delta) - Q_0(\delta)| = o_p(1)$. Thus,

$$\begin{aligned}
\mathbb{P}(Q_0(\hat{\delta}) \geq \kappa) &\leq \mathbb{P}\Big(\sup_{\delta \in \mathcal{D}} |Q_n(\delta) - Q_0(\delta)| > \frac{\kappa}{2}\Big) + o(1) \\
&= o(1).
\end{aligned}$$

Therefore, from $(B.3)$, $\mathbb{P}(||\hat{\delta} - \delta_1|| > \epsilon) \to 0$ as $n \to \infty$, i.e. $\hat{\delta}$ is consistent for δ_1.

Consistency of $\hat{\gamma}$ follows by identical arguments, noting $\hat{\lambda} \xrightarrow{p} 0$, $\hat{\eta} \xrightarrow{p} \eta_0$ and $\hat{\delta} \xrightarrow{p} \delta_1$. □

C Further Assumptions for Semiparametric Estimation

Further restrictions than those stated in the main text are needed for consistency and asymptotic normality of semiparametric estimation in moment condition models. In particular, these additional assumptions ensure convergence properties with plug-in nonparametric estimation. These assumptions relate to Assumptions 5.1-5.4 and 5.6 of Newey [25], pp.1364-7.

Assumption C.1. (i) *There exists* $b_{R_1}(z) \geq 0$ *such that* $||G(\eta) - G - G_1\delta_0' \frac{\partial t}{\partial \eta}(\eta - \eta_0)|| \leq$

$b_{R_1}(z)||\eta-\eta_0||^2$; **(ii)** *there exists* $b_{R_2}(z)\geq 0$ *such that* $||(dG(\eta)^{-1}-1)t(\eta)-(dG^{-1}-1)t+[dG^{-2}G_1t\delta_0'\frac{\partial t}{\partial\eta}-(dG^{-1}-1)\frac{\partial t}{\partial\eta}](\eta-\eta_0)||\leq b_{R_2}(z)||\eta-\eta_0||^2$ *and* $\mathbb{E}[b_{R_2}(z)]\sqrt{n}||\hat{\eta}-\eta_0||^2\xrightarrow{p}0$; **(iii)** *for* $b_1(z):=[\kappa^{-3}b_\psi(z)\kappa_1||\delta_0||^2b_{\partial t}(x)^2+\kappa^{-3}b_\psi(z)b_{R_1}(z)^2+\kappa^{-2}b_\psi(z)b_{R_1}(z)]$, *where* $b_\psi(z)$ *is defined in Assumption 2.1 and* $b_t(x)$, *and* $\mathbb{E}[b_1(z)]\sqrt{n}||\hat{\eta}-\eta_0||^2\xrightarrow{p}0$.

Assumption C.1 states some smoothness restrictions on the functions of η_0 that can be included for over-fitting the propensity score, as well as some boundedness conditions required to control the rate of convergence when first-step nonparametric estimation is involved.

Assumption 5.1 of Newey [25] can be verified as follows. Note that

$$\begin{aligned}\frac{d\psi}{G(\eta)}-\frac{\partial\psi}{G}+\frac{dG_1}{G^2}\psi\delta_0'\frac{\partial t}{\partial\eta}(\eta-\eta_0) &= \frac{d(G-G(\eta))G}{G(\eta)G^2}\psi+\frac{dG(\eta)G_1}{G(\eta)G^2}\psi\delta_0'\frac{\partial t}{\partial\eta}(\eta-\eta_0)\\ &= \frac{d(G-G(\eta))G}{G(\eta)G^2}\psi+\frac{dG(\eta)}{G(\eta)G^2}\psi\Big[G_1\delta_0'\frac{\partial t}{\partial\eta}(\eta-\eta_0)+[G-G(\eta)]\Big]\\ &\quad -\frac{dG(\eta)(G-G(\eta))\psi}{G(\eta)G^2}\\ &= \frac{d\psi(G-G(\eta))^2}{G(\eta)G^2}+\frac{dG(\eta)\psi}{G(\eta)G^2}\Big[G_1\delta_0'\frac{\partial t}{\partial\eta}(\eta-\eta_0)+[G-G(\eta)]\Big].\end{aligned}$$

Thus, by CS and using Assumptions 2.1-2.4, 3.1, 3.2 and C.1(i),

$$\begin{aligned}\Big|\Big|\frac{d\psi}{G(\eta)}-\frac{\partial\psi}{G}+\frac{dG_1}{G^2}\psi\delta_0'\frac{\partial t}{\partial\eta}(\eta-\eta_0)\Big|\Big| &\leq \Big|\Big|\frac{\psi}{G(\eta)G^2}\Big|\Big|\cdot||G-G(\eta)||^2+\Big|\Big|\frac{G(\eta)}{G(\eta)G^2}\Big|\Big|\cdot||\psi||\cdot||b_{R_1}(z)||\\ &\leq \Big|\Big|\frac{\psi}{G(\eta)G^2}\Big|\Big|\Big(\Big|\Big|G_1\delta_0'\frac{\partial t}{\partial\eta}(\eta-\eta_0)\Big|\Big|^2+||b_{R_1}(z)||^2\Big)\\ &\quad +\Big|\Big|\frac{G(\eta)}{G(\eta)G^2}\Big|\Big|\cdot||\psi||\cdot||b_{R_1}(z)||\\ &\leq [\kappa^{-3}b_\psi(z)\kappa_1||\delta_0||^2b_{\partial t}(x)^2+\kappa^{-3}b_\psi(z)b_{R_1}(z)^2+\kappa^{-2}b_\psi(z)b_{R_1}(z)]\\ &\quad \times||\eta-\eta_0||^2,\end{aligned}$$

for $||\eta-\eta_0||<1$. Thus, Assumption C.1(i) and (iii) verifies Assumption 5.1 of Newey [25] for the moment function describing γ_0. The analogous condition for the function $(dG(\eta)^{-1}-1)t(\eta)$ follows from Assumption C.1(ii).

Assumption C.2. $n^{-\frac{1}{2}}\sum_{i=1}^n d_iG_i^{-2}G_{1i}\psi_i\delta_0'\frac{\partial t_i}{\partial\eta}(\hat{\eta}-\eta_0)\xrightarrow{p}n^{-\frac{1}{2}}\sum_{i=1}^n\mathbb{E}[G_1G^{-1}\psi\delta_0'\frac{\partial t}{\partial\eta}|x_{2i}](z_{2i}-\eta_0(x_{2i}))$ *and* $n^{-\frac{1}{2}}\sum_{i=1}^n d_iG_i^{-2}G_{1i}t_i\delta_0'\frac{\partial t_i}{\partial\eta}(\hat{\eta}-\eta_0)\xrightarrow{p}n^{-\frac{1}{2}}\sum_{i=1}^n\mathbb{E}[G_1G^{-1}t\delta_0'\frac{\partial t}{\partial\eta}|x_{2i}](z_{2i}-\eta_0(x_{2i}))$.

Assumption C.2 implies Assumptions 5.2 and 5.3 of Newey ([25], p.1365), the stochastic equicontinuity conditions that relate to the adjustment term that accounts for first-stage nonparametric estimation of η_0. If η_0 is estimated by series approximation, Section 6 of Newey [25] provides sufficient conditions for verifying this condition.

Assumption C.3. *There exists* $\epsilon>0$, $\tilde{b}_1(z)\geq 0$ *and* $\tilde{b}_2(z)\geq 0$ *such that for all* $\delta\in\mathcal{D}$ *and*

$\gamma \in \Gamma$ **(i)** $||dG(t(\eta)'\delta)^{-1}\psi(\gamma) - dG(t(\eta_0)'\delta)^{-1}\psi(\gamma)|| \leq \tilde{b}_1(z)$**; (ii)** $||(dG(t(\eta)'\delta)^{-1} - 1)t(\eta) - (dG(t(\eta_0)'\delta)^{-1} - 1)t(\eta_0)|| \leq \tilde{b}_2(z)$.

Assumption C.3 re-states Assumption 5.4(ii) of Newey ([25], p.1367). For the verification of Assumption 5.4(i) of Newey [25] note that for any $\delta \in \mathcal{D}$ and $\psi \in \Gamma$,

$$\begin{aligned} \Big|\Big|\frac{d}{G(t(\eta)'\delta)}\psi(z,\gamma)\Big|\Big| &\leq \kappa^{-1}b_\psi(z) \\ \Big|\Big|\Big(\frac{d}{G(t(\eta)'\delta)} - 1\Big)t(\eta)\Big|\Big| &\leq (\kappa^{-1}+1)b_t(x), \end{aligned}$$

where $\mathbb{E}[b_\psi(z)] < \infty$ and $\mathbb{E}[b_t(z)] < \infty$ by Assumptions 2.1 and 3.2. This, together with Assumption C.3, allows uniform convergence of $\delta \in \mathcal{D}$ and $\gamma \in \Gamma$, that is,

$$\begin{aligned} \sup_{\delta\in\mathcal{D}}\Big|\Big|\frac{1}{n}\sum_{i=1}^n\Big(\frac{d_i}{G(t_i(\hat{\eta})'\delta)} - 1\Big)t_i(\hat{\eta}) - \mathbb{E}\Big[\Big(\frac{d}{G(t'\delta)} - 1\Big)t\Big]\Big|\Big| &\xrightarrow{p} 0 \\ \sup_{\delta\in\mathcal{D},\,\gamma\in\Gamma}\Big|\Big|\frac{1}{n}\sum_{i=1}^n\frac{d_i\psi_i(\gamma)}{G(t_i(\hat{\eta})'\delta)} - \mathbb{E}\Big[\frac{d\psi(\gamma)}{G(t'\delta)}\Big]\Big|\Big| &\xrightarrow{p} 0. \end{aligned}$$

Finally, additional assumptions are needed to verify Assumption 5.6 of Newey ([25], p.1367). To verify Assumption 5.6(iv) of Newey [25], note that, by Assumption 3.2,

$$\begin{aligned} \mathbb{E}\Big[\Big|\Big|\Big(\frac{d}{G} - 1\Big)t\Big|\Big|^2\Big] &< (\kappa^{-1}+1)^2\mathbb{E}[b_t(x)^2] < \infty \\ \mathbb{E}\Big[\Big|\Big|\frac{d}{G}\psi\Big|\Big|^2\Big] &< \kappa^{-2}\mathbb{E}[b_\psi(z)^2] < \infty. \end{aligned}$$

Assumption C.4. (i) *Assumption C.3(i) is satisfied with* $dG(t(\eta)'\delta)^{-1}\psi(\gamma)$ *there equal to each row of* $dG(t(\eta)'\delta)^{-1}\Psi(\gamma)$**; (ii)** *Assumption C.3(ii) is satisfied with* $dG(t(\eta)'\delta)^{-1} - 1)t(\eta)$ *there equal to each row of* $dG(\eta)^{-2}G_1(\eta)t(\eta)t(\eta)'$.

Assumption C.4 re-states Assumption 5.6(v) of Newey [25], which allows uniform convergence of the Jacobian. This is useful to construct consistent estimators of the asymptotic variances of $\hat{\delta}$ and $\hat{\gamma}$.

Ph.D. Thesis: Chapter Two

Robust Estimation Using Auxiliary Semiparametric Information

Ashish Patel*
University of Cambridge

Abstract

This paper concerns semiparametric moment condition models where the parameter of interest is described by one set of moment restrictions, while nuisance functions are identified from another set of moment restrictions. A two-step generalised empirical likelihood-weighted estimator is proposed that guarantees an efficiency gain arising from exploiting auxiliary moment restrictions that may involve nonparametric components. To achieve this, moment restrictions generally need to be adjusted to account for first-stage nuisance estimation of nonparametric components; see Chernozhukov et al. [12]. This paper adapts and generalises the approach of Hellerstein and Imbens [21] and Bravo [6] to models specified by semiparametric moment conditions with estimated nuisance functions. A leading example considered here is a semiparametric missing data model. An efficiency gain under misspecification is possible as compared with estimation based on the efficient influence function of the semiparametric missing data model studied in Graham [17], while also preserving double robustness properties.

Keywords: Generalised Empirical Likelihood, Semiparametric Moment Conditions, Local Robustness, Double Robustness, Missing Data

*I thank Richard Smith for detailed comments and helpful discussions. I thank Oliver Linton and Shaun Seaman for helpful comments. I gratefully acknowledge financial support received from an ESRC Studentship Award.

1 Introduction

This paper considers robust estimation of a finite-dimensional parameter that is described by one set of moment restrictions, when there also exist extra moment restrictions describing and identifying nuisance functions. In many applications, such a moment conditions set-up arises naturally. For example, many models in the missing data and treatment effects literature fit into this general framework. Following Graham [17] the moment restrictions that describe the parameter of interest are described as the *identifying restrictions*, whereas the moment restrictions describing only the nuisance functions are the *auxiliary restrictions*. It is well known that, for efficiency considerations, the information from both identifying and auxiliary restrictions should be used, see, for example, Prokhorov and Schmidt [31].

Two-step efficient generalised method of moments (GMM) and generalised empirical likelihood (GEL) are commonly used to estimate moment condition models. Under correct model specification, both classes of estimators are asymptotically first-order equivalent. However, GEL has been shown to have advantages both from finite-sample (Hansen et al. [20], Newey et al. [27]) and higher-order asymptotic considerations (Newey and Smith [28]).

When unconditional moment restrictions contain unknown functions, Newey [25] and Chen et al. [11] provide regularity conditions for $\sqrt{n}$-consistency and asymptotic normality of GMM with plug-in nonparametric estimates. Hjört et al. [22] provide regularity conditions for empirical likelihood estimators with the same properties as GMM. One-step methods exist that jointly estimate the nonparametric functions while also obtaining semiparametrically efficient estimators of finite dimensional parameters. The sieve-GMM estimator of Ai and Chen [2] uses approximating functions for models characterised by conditional moment restrictions with endogenous nuisance functions. Otsu [29] proposes an analogous estimator for empirical likelihood.

One-step estimators implicitly take the optimal linear combination of moment functions for the purposes of efficiency. This approach relies on all moment restrictions being correctly specified. However, when auxiliary moment restrictions are misspecified the properties of one-step efficient estimators may be a concern. As a result, there has been considerable interest in the development of estimators robust with respect to nuisance functions. The concept of *double robustness* originated in the study of missing data models, see Robins et al. [34] and Scharfstein et al. [35]. Doubly robust estimators are consistent if at least one of two nuisance functions are consistently estimated; this has proved to be an attractive property in practice, particularly in biostatistics and econometrics.

Doubly robust methods have been also shown to be valuable when nuisance functions are nonparametrically-estimated. Frölich et al. [16] illustrate the finite-sample advantages of nonparametrically-estimated policy evaluation parameters based on doubly robust estimating equations. Firpo and Rothe [15] show doubly robust moment conditions may permit

larger smoothing biases of nonparametric estimators in the first stage, while maintaining $\sqrt{n}$-consistency for estimators of a finite dimensional parameter of interest. The related property of *local robustness* has been characterised by Chernozhukov et al. [12] and provides a more general basis for estimation that is desensitised with respect to first stage estimation of nuisance functions.

Robustness properties give practitioners more confidence in proposing working models or relationships that can be characterised by a set of auxiliary moment restrictions. In this paper, the trade-off between efficiency and robustness of using auxiliary restrictions is examined. A two-step GEL-weighted estimator is proposed whereby information from the auxiliary restrictions guarantees efficiency gains for the parameter of interest while also aiding robustness properties. As such, our approach offers a middle ground for efficiency-robustness considerations. In the first step, a locally robust version of the auxiliary moment restrictions are estimated by GEL. In the second step, a simple method of moments estimator is re-weighted by sample weights arising from first-step GEL estimation. The extent of the efficiency gain depends on the correlation between the identifying and the auxiliary restrictions.

Furthermore, this estimation procedure offers more control than one-step efficient estimators over how moment conditions are combined, thereby allowing for the preservation of robustness properties when auxiliary information is included. A leading example is the semiparametric missing data model of Graham [17]. For doubly robust estimation of missing data models, nuisance estimation of the *propensity score*, the probability that data are missing conditional on observables, and the *conditional outcome function*, the conditional expectation of the moment function given observables, is required. It is shown that this approach guarantees an efficiency gain under misspecification of either the propensity score or conditional outcome model over estimation based on the efficient influence function of the semiparametric missing data model. Indeed, this approach preserves double robustness properties under less stringent requirements than estimation based on the efficient influence function.

This paper is organised as follows. Section 2 introduces the semiparametric moment conditions set-up and GEL estimation. Section 3 presents the two-step GEL-weighted estimator and Section 4 presents its asymptotic properties. Section 5 discusses the semiparametric missing data model and applies the results of this paper to that context. Section 6 illustrates the use of the estimation method by simulation. Section 7 concludes.

The following abbreviations are used. $\xrightarrow{p}$: converges in probability to, $\xrightarrow{d}$: converges in distribution to, T : the triangle inequality, CS: the Cauchy-Schwarz inequality, M: the Markov inequality, UWL: the uniform weak law of large numbers (for example, Lemma 2.4 of Newey and McFadden [26]), CLT: a central limit theorem for i.i.d. random variables, WLLN: the weak law of large numbers, LIE: the law of iterated expectations, w.p.1: with probability one, w.p.a.1: with probability approaching one, LHS: left hand side, RHS: right hand side, and $||.||$ is the Euclidean norm.

2 Semiparametric Moment Restrictions and GEL

2.1 Semiparametric moment restrictions

Let z denote a d_z-vector of random variables taking values in $\mathcal{Z}$. Also, let $g_1(z, \theta_1, h_1)$ be a d_{g_1}-vector of known continuous functions of the data vector z, a d_{θ_1}-dimensional unknown vector $\theta_1 \in \Theta_1$ where $\Theta_1 \subset \mathbb{R}^{d_{\Theta_1}}$ is compact, and unknown functions $h_1 : \mathcal{X} \times \Theta_1 \to \mathbb{R}^{d_{h_1}}$, where $x \in \mathcal{X}$ is a d_x-vector contained in $z \in \mathcal{Z}$. Let $g_2(z, \theta_2, h_2)$ be a d_{g_2}-vector of known continuous functions of z, a d_{θ_2}-dimensional unknown vector $\theta_2 \in \Theta_2$ where $\Theta_2 \subset \mathbb{R}^{d_{\theta_2}}$ is compact, and unknown functions $h_2 : \mathcal{X} \times \Theta_2 \to \mathbb{R}^{d_{h_2}}$.

The parameter θ_1 is of inferential interest, whereas θ_2, $h_1(x, \theta_1)$ and $h_2(x, \theta_2)$ can be regarded as nuisance parameters, with $h_1(x, \theta_1)$ and $h_2(x, \theta_2)$ being infinite-dimensional.

The moment condition model comprises identifying and auxiliary restrictions such that

$$\text{identifying restriction: } \mathbb{E}[g_1(z, \theta_{10}, h_{10})] = 0 \tag{2.1}$$

$$\text{auxiliary restriction: } \mathbb{E}[g_2(z, \theta_{20}, h_{20})] = 0 \tag{2.2}$$

where expectation is taken with respect to the distribution of z. It is assumed that (2.1) and (2.2) hold uniquely at $\theta_{10} \in \Theta_1$ and $\theta_{20} \in \Theta_2$, respectively. h_{10} and h_{20} denote the nuisance functions evaluated at the true values of θ_{10} and θ_{20}, respectively.

To identify θ_{10}, $d_{g_1} \geq d_{\theta_1}$ is required. For ease of exposition, however, the just-identified case $d_{g_1} = d_{\theta_1}$ is considered, although the approach can be easily extended to the over-identified case. Similarly, for the identification of θ_{20}, $d_{g_2} \geq d_{\theta_2}$ is also required; the over-identified case $d_{g_2} > d_{\theta_2}$ is considered here which describes many works where the auxiliary model is a conditional probability or comprises auxiliary information, see, for example, Hellerstein and Imbens [21] and Bravo [6].

Example 2.1 (**average treatment effects**). Let $z = (y, x, d)'$, where y is a scalar denoting the outcome. Let d be a binary random variable indicating whether or not an individual has been treated ($d = 1$ if treated). Let x be a d_x-vector of individual characteristics, and x_2 be a subset of characteristics of x. The propensity score $p_0(x) = \mathbb{P}(d = 1|x)$ is modelled by a semiparametric model $p(x) = r(x, \theta_2, h_2(x_2))$, where θ_2 is a d_{θ_2}-dimensional unknown parameter and h_2 is an unknown function of x_2. Finally let $t(x)$ be a d_t-dimensional vector of known, independent functions of x such that $d_t > d_{\theta_2}$. A semiparametric model for the average treatment effect θ_{10} is given by (2.1) and (2.2) with

$$\begin{aligned} g_1(z, \theta_1, \theta_2, h_2) &= \frac{yd}{r(x, \theta_2, h_2(x_2))} - \theta_1 \\ g_2(z, \theta_2, h_2) &= \Big(\frac{d}{r(x, \theta_2, h_2(x_2))} - 1\Big) t(x). \end{aligned}$$

Consistent estimation of θ_1 relies on the missing at random assumption $y \perp d|x$, that is, the outcome is independent of treatment given covariates. See Graham et al. [18] for a similar specification of the moment condition describing the propensity score model.

2.2 GEL estimation

Given a sample $\{z_i\}_{i=1}^n$, and given a nonparametric estimator $\hat{h}_2$ of h_{20}, the GEL estimation method for θ_2 is as follows. Let

$$\hat{P}_n(\theta_2, \hat{h}_2, \lambda) = \frac{1}{n}\sum_{i=1}^{n}[\rho(\lambda' g_2(z_i, \theta_2, \hat{h}_2)) - \rho_0]$$

where the function $\rho(.)$ is concave on its domain $\mathcal{V}$, an open interval containing zero, with derivatives $\rho_j(v) = \partial^j \rho(v)/dv^j$, $\rho_j(0) = \rho_j$, $j = 0, 1, ...$, normalised without loss of generality as $\rho_1 = \rho_2 = -1$. The GEL estimator of θ_2 is given by

$$\hat{\theta}_2 = \arg\min_{\theta_2 \in \Theta_2} \sup_{\lambda \in \Lambda_n} \hat{P}_n(\theta_2, \hat{h}_2, \lambda)$$

where $\Lambda_n = \{\lambda : ||\lambda|| \leq n^{-\zeta}\}$ for some $\frac{1}{\alpha} < \zeta < \frac{1}{2}$ and $\alpha > 2$. For any $\theta_2 \in \Theta_2$, an estimator of the d_{g_2}-vector of auxiliary parameters is given by $\hat{\lambda}(\theta_2, \hat{h}_2) = \text{argmax}_{\lambda \in \Lambda_n} \hat{P}_n(\theta_2, \hat{h}_2, \lambda(\theta_2, \hat{h}_2))$. Let $\hat{\lambda} = \hat{\lambda}(\hat{\theta}_2, \hat{h}_2)$. The first-order condition for λ imposes the sample moment constraint $\sum_{i=1}^n \hat{\pi}_i g_2(z_i, \hat{\theta}_2, \hat{h}_2) = 0$ where the GEL implied probabilities are $\{\hat{\pi}_i\}_{i=1}^n$ are given by

$$\hat{\pi}_i = \frac{\rho_1\big(\hat{\lambda}' g_2(z_i, \hat{\theta}_2, \hat{h}_2)\big)}{\sum_{j=1}^n \rho_1\big(\hat{\lambda}' g_2(z_j, \hat{\theta}_2, \hat{h}_2)\big)}, \quad (i = 1, ..., n). \tag{2.3}$$

The auxiliary parameter λ may be interpreted as the Lagrange multiplier associated with the sample moment constraint $\sum_{i=1}^n \hat{\pi}_i g_2(z_i, \hat{\theta}_2, \hat{h}_2) = 0$. Special cases include:

- empirical likelihood (EL): $\rho(v) = \ln(1 - v)$ and $\mathcal{V} = (-\infty, 1)$, resulting in implied probabilities $\hat{\pi}_i^{EL} = n^{-1}\big(1 + \hat{\lambda}' g_2(z_i, \hat{\theta}_2, \hat{h}_2)\big)^{-1}$ $(i = 1, ..., n)$.
- exponential tilting (ET): $\rho(v) = 1 - \exp(v)$, resulting in implied probabilities $\hat{\pi}_i^{ET} = \exp\big(\hat{\lambda}' g_2(z_i, \hat{\theta}_2, \hat{h}_2)\big) / \sum_{j=1}^n \exp\big(\hat{\lambda}' g_2(z_j, \hat{\theta}_2, \hat{h}_2)\big)$ $(i = 1, ..., n)$.

GEL methods have been shown to possess some advantages over GMM. Newey et al. [27] present Monte Carlo results indicating smaller bias properties of GEL estimators in small samples compared to GMM. Furthermore, while the first-order asymptotic properties of GEL and GMM estimators are identical, Newey and Smith [28] show the second-order asymptotic bias for GEL estimators comprises of fewer components compared to that of GMM, with the EL estimator being the best by this measure. On a practical note, in contrast to two-step efficient GMM estimation, GEL methods do not require estimation of the Jacobian or covariance matrix $\Omega_2 = \mathbb{E}[g_2(z, \theta_{20}, h_{20}) g_2(z, \theta_{20}, h_{20})']$.

2.3 GEL implied probabilities

The implied probabilities of an estimator reveal how much weight GEL estimation places on each observation. When a model is correctly specified, all observations should be equally representative and thus the implied probabilities should hover close to n^{-1}. Back and Brown [5] derived the implied probabilities associated with GMM estimation. The implied probabilities for GEL estimators are given in Smith [37] and Newey and Smith [28].

Several developments based on implied probabilities have been made for moment condition model estimation. Schennach [36] considers a hybrid estimator involving the attractive properties of both ET and EL estimation that improves the behaviour of implied probabilities under misspecification. Antoine et al. [4] study an estimator that minimises the distance of a distribution function implied by the moment conditions to the empirical distribution in χ^2 distance. While the implied probabilities of some GEL estimators are guaranteed to be positive in large samples, Antoine et al. [4] propose a simple correction that restores positivity of implied probabilities in finite samples.

Estimation procedures which make explicit use of GEL implied probabilities display particular advantages. Brown and Newey [8] show that a sample moment condition using EL implied probability weights results in a semiparametrically efficient estimator for the associated population moment. Brown and Newey [9] use implied probabilities to devise an efficient bootstrap procedure for inference.

In closely related work, for missing data models, Chen et al. [10] discuss how one-step estimation of all moment restrictions stacked together may not be efficient when some of the moment restrictions are misspecified. By re-weighting those estimating equations with EL implied probabilities for the moment conditions describing the conditional expectation function, efficiency gains are possible.

Using a similar approach, Bravo [6] considers re-weighting an estimating equation for m-estimators by GEL implied probabilities of a known moment condition. It is shown that when no nuisance estimation is involved, such a procedure efficiently incorporates the information of the moment condition so that efficiency gains are guaranteed. Furthermore, Bravo [6] shows a two-step procedure may have higher-order benefits.

Existing results which make explicit use of GEL implied probabilities for re-weighing estimating equations suggest the procedure guarantees efficiency gains from using information from auxiliary restrictions while preserving robustness. Issues of efficiency and robustness are explored in subsequent sections.

3 Two-step GEL Estimation

3.1 Locally robust moment conditions

Chernozhukov et al. [12] introduce locally robust moment conditions that are designed to be insensitive to nuisance function estimates. Local robustness is defined in terms of the limit of functions, $h_1(F)$, say, which vary with the true distribution F. Consider families of regular parametric models F_τ indexed by a vector of parameters τ such that the true distribution of z is F_0 with score $S(z)$. Without loss of generality, the focus is on the moment function $g_1(z, \theta_1, h_1)$.

Definition. The moment function $g_1(z, \theta_1, h_1)$ is said to be *locally robust* if and only if for all regular parametric models $\partial \mathbb{E}[g_1(z, \theta_{10}, h_1(F_\tau))]/\partial \tau\big|_{\tau=0} = 0$.

That is, the moment conditions are insensitive to movements in $h_1(F)$ away from the true distribution F_0.

The procedure for constructing locally robust moment functions involves calculating an adjustment term $\phi_1(z, \theta_1, h_1)$ that accounts for first-step estimation of the function h_{10}, that satisfies

$$\begin{aligned} \partial \mathbb{E}[g_1(z, \theta_{10}, h_1(F_\tau))]/\partial \tau &= \mathbb{E}[\phi_1(z, \theta_{10}, h_{10}) S(z)] \\ \mathbb{E}[\phi_1(z, \theta_{10}, h_{10})] &= 0, \end{aligned}$$

where the details for the computation of $\phi_1(z, \theta_1, h_1)$ are discussed in Chernozhukov et al. [12], pp.6-11. Although not explicitly stated, other nonparametric functions may need to be estimated for this orthogonalisation step which means the function h_1 in $\phi_1(z, \theta_1, h_1)$ may comprise additional nuisance functions not included in the original moment indicator vector $g_1(z, \theta_1, h_1)$. After computing $\phi_1(z, \theta_1, h_1)$, see Proposition 2, p.11, of Chernozhukov et al. [12], the adjusted moment function $\psi_1(z, \theta_1, h_1) = g_1(z, \theta_1, h_1) + \phi_1(z, \theta_1, h_1)$ is locally robust. Under regularity conditions stated there, for a suitable nonparametric estimator $\hat{h}_1$,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_1(z_i, \theta_{10}, \hat{h}_1(x_i)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_1(z_i, \theta_{10}, h_{10}(x_i)) + o_p(1) \tag{3.1}$$

Therefore, if estimation of θ_{10} is based on the locally robust moment condition $\mathbb{E}[\psi_1(z, \theta_{10}, h_{10})] = 0$, the asymptotic variance of $\hat{\theta}_1$ is invariant to nonparametric estimation of $h_{10}(x)$.

A well-known example of such a locally robust moment condition concerns estimation of the average treatment effect. With the same notation used in Example 2.1, the original moment function $g_1(z, \theta_1, p) = \theta_1 - yd/p(x)$ describes the average treatment effect (when outcomes y in absence of treatment are zero), where $p_0(x)$ is the propensity score. The adjustment term, see Hahn [19], is given by $\phi_1(z, p, \gamma) = ((d/p(x)) - 1)\gamma(x)$ where $\gamma_0(x)$ is the average outcome

conditional on observables $\mathbb{E}[y|x]$. Computing the adjustment term involves estimating two nuisance functions $p_0(x)$ and $\gamma_0(x)$, though neither play a role in the asymptotic variance of $\hat{\theta}_1$ if estimated using the locally robust moment function $\mathbb{E}[\psi_1(z, \theta_{10}, p_0, \gamma_0)] = 0$ where $\psi_1(z, \theta_1, p, \gamma) = g_1(z, \theta_1, p) + \phi_1(z, p, \gamma)$.

3.2 Two-step GEL-weighted estimation

Consider two-step GEL-weighted estimation based on the moment conditions set-up (2.1), (2.2). The first-step involves GEL estimation of a locally robust version of (2.2).

3.2.1 Step 1: Obtain GEL implied probabilities $\{\hat{\pi}_i\}_{i=1}^n$

For $j = 1, 2$, given the moment function $g_j(z, \theta_j, h_j)$, denote the adjustment term that accounts for plug-in estimation of h_{j0} by $\phi_j(z, \theta_j, h_j)$. Although not explicit it is understood that h_j may also contain other nonparametric functions needed to calculate the adjustment term. The locally robust moment function is constructed as $\psi_j(z, \theta_j, h_j) = g_j(z, \theta_j, h_j) + \phi_j(z, \theta_j, h_j)$ $(j = 1, 2)$.

Given a nonparametric estimate $\hat{h}_2$ of h_{20}, θ_{20} is estimated by GEL based on the locally robust moment condition $\mathbb{E}[\psi_2(z, \theta_{20}, h_{20})]$. Given the GEL estimator $\hat{\theta}_2$ and associated Lagrange multiplier $\hat{\lambda}$, the GEL implied probabilities $\{\hat{\pi}_i\}_{i=1}^n$ are defined as in (2.3) with the locally robust moment function ψ_2 replacing g_2.

3.2.2 Step 2: Weighted Method of Moments estimation of θ_{10}

Given the implied probabilities $\{\hat{\pi}\}_{i=1}^n$, and a nonparametric estimator $\hat{h}_1$ for h_{10}, the estimator $\hat{\theta}_1$ for θ_{10} solves the re-weighted moment equations

$$\sum_{i=1}^{n} \hat{\pi}_i \psi_1(z_i, \hat{\theta}_1, \hat{h}_1(x_i)) = 0. \tag{3.2}$$

3.3 One-step versus Two-step Estimation

One-step estimation with sieve-GMM/GEL estimation (Ai and Chen [2], Otsu [29]) leads to inherently efficient estimation when moment restrictions are correctly specified. However, there are reasons why two-step approaches remain popular in practice. Ackerberg et al. [1] shows how locally robust-type corrections can lead to fully efficient estimation in a large class of semiparametric models with plug-in nonparametric estimation.

Furthermore, the two-step procedure has computational advantages. As argued by Ackerberg et al. [1], the joint approach requires a large-dimensional non-linear search over θ (finite-dimensional components) and h (infinite-dimensional components) simultaneously. Thus, in

terms of computational time and reliability, sequential procedures involving first-step estimation of h and second step estimation of θ may have advantages. Sequential estimation of moment restrictions also imposes less stringent restrictions on boundedness of moments for large sample results. This is briefly discussed in the next section.

The main disadvantage of two-step procedures is that the moment restrictions are not simultaneously exploited. The two-step procedure considered in this paper mitigates this issue by incorporating information from the auxiliary restrictions into the identifying moment restrictions to preserve efficiency gains. Note that although this approach may not lead to the optimal linear combination of moment restrictions for semiparametric efficiency, more control is possible to preserve known robustness properties.

For example, Chen et al. [10] show that estimation based on the efficient influence function of the missing data model considered by Robins et al. [34] may not necessarily be the best way to use auxiliary restrictions under forms of misspecification. This result is extended in Section 5 to show how a similar GEL-weighted approach can preserve double robustness properties in a large class of semiparametric missing data models.

The GEL-weighted two-step approach may also have benefits in terms of higher-order bias. One of the components of higher order asymptotic bias of GEL estimators, as characterised by Newey and Smith [28], is linked with the degree of over-identification of moment condition models. Bravo [6] applies these results to show that in the case where the auxiliary restrictions involve no nuisance estimation, a GEL-weighted approach is not only semiparametrically efficient but also higher order efficient relative to one-step estimation when all moment conditions are exploited jointly.

Finally, Hellerstein and Imbens [21] shows a similar two-step estimation strategy may alleviate selection bias. More specifically, suppose an available sample $\{z\}_{i=1}^{n}$ is taken from a sample distribution $f_s(z)$ which is different from the population (or target) distribution $f_t(z)$. The parameter of interest θ_{10} is described by a moment condition $\mathbb{E}_t[g_1(z,\theta_{10})] = 0$ and an auxiliary moment restriction $\mathbb{E}_t[g_2(z)] = 0$ holds, where $\mathbb{E}_t[.]$ denotes the expectation taken over the target distribution $f_t(z)$. Hellerstein and Imbens ([21], Theorem 2 and Section IV, pp.4-6) shows that, in the case with no nonparametric or nuisance parameters, so that $\psi_1(z,\theta_1,h_1) = g_1(z,\theta_1)$ and $\psi_2(z,\theta_2,h_2) = g_2(z)$, an estimator based on (3.2) reduces sample selection bias.

In particular, the estimator converges to the probability limit of a method of moments estimator of $\mathbb{E}[g_1(z,\theta_0)] = 0$ using a random sample from an artificial population with distribution $f_{st}(z)$, where this artifical distribution has the interpretation of being the closest to $f_s(z)$ in an empirical likelihood sense such that $\mathbb{E}_{st}[g_1(z,\theta_0)] = \mathbb{E}_t[g_1(z,\theta_0)] = 0$. That is, it ensures the artificial distribution shares the auxiliary moment restriction with the target distribution which may help reduce bias in estimation.

4 Theoretical Results

4.1 Assumptions

The following assumptions are maintained for the large sample results. Since the identifying and auxiliary restrictions are estimated separately, in theory, less stringent requirements are needed to estimate the identifying moment restrictions. For example, in contrast to one-step estimation where all moment functions are exploited simultaneously, no assumptions are required on cross-derivatives between identifying moments and the auxiliary restrictions. Furthermore, if the parameter of interest is identified only by just-identified restrictions with no nuisance function estimation, only the standard conditions on g_1 (2.1) necessary for method of moments estimation (Newey and McFadden [26], Theorems 2.1 and 3.1) are needed. Hence, the imposition of further boundedness conditions required for joint GMM and GEL may be avoided.

For simplicity of exposition, such distinctions are not made here and the same conditions are imposed on $g_1(z, \theta_1, h_1)$ and $g_2(z, \theta_2, h_2)$. The general conditions are similar to Newey [25] and Chen et al. [11]. The additional assumptions for GEL estimation follow Newey and Smith [28] and Hjört et al. [22]; also see Bravo et al. [7]. For $j \in \{1, 2\}$, let $G_j(z, \theta_j, h_j) = \partial g_j(z, \theta_j, h_j)/\partial \theta_j$, $G_j = \mathbb{E}[G_j(z, \theta_{j0}, h_{j0})]$, and $\Omega_j = \mathbb{E}[g_j(z, \theta_{j0}, h_{j0}) g_j(z, \theta_{j0}, h_{j0})']$.

Assumption. 4.1 *For $j \in \{1, 2\}$,* **(i)** *$\theta_{j0} \in \Theta_j$ is the unique solution to $\mathbb{E}[g_j(z, \theta_j, h_j(\theta_j))] = 0$ and $\theta_{j0} \in int(\Theta_j)$;* **(ii)** *$G_j$ has full column rank d_{θ_j};* **(iii)** *Ω_j is nonsingular;* **(iv)** *there exists a function $D_j(z, h_j)$ linear in h_j, and $c_j(z)$, such that for all $||h_j - h_{j0}||$ small enough, $||g_j(z, \theta_{j0}, h_j) - g_j(z, \theta_{j0}, h_{j0}) - D_j(z, h_j - h_{j0})|| \leq c_j(z)||h_j - h_{j0}||^2$ and $\sum_{i=1}^n [D_j(z_i, \hat{h}_j - h_{j0}) - \int D_j(z, \hat{h}_j - h_{j0}) dF_0]/\sqrt{n} \xrightarrow{p} 0$;* **(v)** *$\mathbb{E}[c_j(z)]\sqrt{n}||\hat{h}_j - h_{j0}||^2 \xrightarrow{p} 0$ and $\mathbb{E}[c_j(z)] < \infty$;* **(vi)** *$\rho(v)$ is three times continuously differentiable in a neighborhood of zero.*

Assumption 4.1 is sufficient for consistent GMM and GEL estimation of θ_{10} and θ_{20}. Assumption 4.1(iv) requires the moment functions to be sufficiently smooth in nonparametric components. In the parametric case, if h_j is a finite-dimensional parameter, $D_j(z, h_j - h_{j0})$ corresponds to the term $[\partial g_j(z, \theta_{j0}, h_{j0})/\partial h_j](\hat{h}_j - h_{j0})$. Assumption 4.1(iv) also contains a stochastic equicontinuity condition which is standard in the literature, and can be verified using conditions given in Andrews [3]. Assumption 4.1(v) is a restriction to guarantee the remainder term converges fast enough; $\hat{h}_j$ is required to converge to h_{j0} at a rate faster than $n^{-\frac{1}{4}}$ and is satisfied for commonly used nonparametric estimators (series, kernels etc.) if the function h_j is sufficiently smooth. Assumption 4.1(vi) restricts the class of concave functions chosen for GEL estimation.

Assumption. 4.2 *For $j \in \{1, 2\}$, there is a function $\phi_j(z, \theta_j, h_j)$ such that $\mathbb{E}[\phi_j(z, \theta_{j0}, h_{j0})] = 0$, $\mathbb{E}[||\phi_j(z, \theta_{j0}, h_{j0})||^\alpha] < \infty$, $\sqrt{n} \int D_j(z, \hat{h}_j - h_{j0}) dF_0 - \sum_{i=1}^n \phi_j(z, \theta_{j0}, h_{j0})/\sqrt{n}$ and $\sum_{i=1}^n ||\phi_j(z_i, \hat{\theta}_j, \hat{h}_j) - \phi_j(z_i, \theta_{j0}, h_{j0})||^2/n \xrightarrow{p} 0$.*

Assumptions 4.1 and 4.2 imply $\sum_{i=1}^n g_j(z_i, \theta_{j0}, \hat{h}_j)/\sqrt{n}$ is asymptotically normal. Section 3.1 discusses methods to calculate $\phi_j(z, \theta_j, h_j)$ such that this assumption holds. Also see Newey [25], pp. 1366-7, for further discussion on pathwise derivative calculations.

Assumption. 4.3 *For $j \in \{1, 2\}$, there exist $\epsilon > 0$, $||h_j||$, $b_j(z)$, $\tilde{b}_j(z)$, $d_j(z)$ and $\tilde{d}_j(z) > 0$ such that for all $\theta_j \in \Theta_j$ and $||h_j - h_{j0}|| < \epsilon$,* **(i)** *$g_j(z, \theta_j, h_j)$ is continuous at θ_j w.p.1, $||g_j(z, \theta_j, h_j)|| \leq d_j(z)$ and $\mathbb{E}[d_j(z)^\alpha] < \infty$ for some $\alpha > 2$;* **(ii)** *$||G_j(z, \theta_j, h_j)|| \leq \tilde{d}_j(z)$ and $\mathbb{E}[\tilde{d}_j(z)^2] < \infty$;* **(iii)** *$||g_j(z, \theta_j, h_j) - g_j(z, \theta_{j0}, h_{j0})|| \leq b_j(z)(||\theta_j - \theta_{j0}|| + ||h_j - h_{j0}||)$ and $\mathbb{E}[b_j(z)^2] < \infty$;* **(iv)** *$||G_j(z, \theta_j, h_j) - G_j(z, \theta_{j0}, h_{j0})|| \leq \tilde{b}_j(z)(||\theta_j - \theta_{j0}|| + ||h_j - h_{j0}||)$ and $\mathbb{E}[\tilde{b}_j(z)] < \infty$;* **(vii)** *there exist $||h_j||$, $\epsilon > 0$ and a neighbourhood $\mathcal{N}_{\theta_j}$ of θ_{j0} such that for all $||h_j - h_{j0}|| < \epsilon$, $g_j(z, \theta_j, h_j)$ is twice continuously differentiable in $\theta_j \in \mathcal{N}_{\theta_j}$.*

Assumptions 4.1-4.3 are sufficient for the asymptotic normality of semiparametric GEL estimation of θ_{20}. Assumption 4.3 guarantees the remainder term from nonparametric estimation of h_{j0} ($j = 1, 2$) is small, and allows for uniform law of large numbers results required for uniform convergence of $\hat{\theta}_j$ to θ_{j0}, that is, $\sup_{\theta_j \in \Theta_j} ||(\sum_{i=1}^n g_j(z_i, \theta_j, \hat{h}_j)/n) - \mathbb{E}[g_j(z, \theta_j, h_{j0})]|| \xrightarrow{p} 0$ ($j = 1, 2$). Assumptions 4.1-4.3 can also be used to establish consistency for components relating to the asymptotic variance of $\hat{\theta}_j$, in particular, G_j and Ω_j ($j = 1, 2$).

Assumption. 4.4 (i) *For $j = 1, 2$, $b_j(z)$, $d_j(z)$, $\tilde{d}_j(z)$ satisfying Assumption 4.3, $\mathbb{E}[b_2(z)d_2(z)] < \infty$, $\mathbb{E}[b_1(z)d_2(z)] < \infty$, $\mathbb{E}[d_2(z)\tilde{d}_2(z)] < \infty$, $\mathbb{E}[b_2(z)d_1(z)] < \infty$, $\mathbb{E}[d_1(z)b_2(z)] < \infty$ and $\mathbb{E}[\tilde{d}_1(z)d_2(z)] < \infty$;* **(ii)** *Assumptions 4.1, 4.2, 4.3 and 4.4(i) hold for the moment function $\psi_j(z, \theta_j, h_j)$ replacing $g_j(z, \theta_j, h_j)$ ($j = 1, 2$).*

Assumption 4.4 states additional boundedness conditions relating to the correlation matrices of the moment functions g_1 and g_2. Assumptions 4-1-4.4 are sufficient for $\hat{\theta}_1$ to be $\sqrt{n}$-consistent and asymptotically normal.

4.2 Asymptotic results

This section presents a large sample result on the efficiency gains from using auxiliary information for the estimator $\hat{\theta}_1$ obtained from (3.2). Define the following matrices.

$$V_j = \mathbb{E}[(g_j(z, \theta_{j0}, h_{j0}) + \phi_j(z, \theta_{j0}, h_{j0}))(g_j(z, \theta_{j0}, h_{j0}) + \phi_j(z, \theta_{j0}, h_{j0}))'], \quad (j = 1, 2)$$

where $\phi_j(z, \theta_{j0}, h_{j0})$ is the adjustment term that embodies the effect of estimating h_{j0} on the moment function g_j. Assumptions 4.1-4.3 satisfy the hypotheses of Lemma 5.1 of Newey [25]. Hence, for a nonparametric estimator $\hat{h}_j$ for h_{j0} converging at a rate faster than $n^{-\frac{1}{4}}$,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n g_j(z_i, \theta_{j0}, \hat{h}_j) \xrightarrow{d} \mathcal{N}(0, V_j).$$

For $j = 1, 2$, let $P = \Omega_2^{-1} - \Omega_2^{-1} G_2 (G_2' \Omega_2^{-1} G_2)^{-1} G_2' \Omega_2^{-1}$, and $B = \mathbb{E}[g_1(z, \theta_{10}, h_{10}) g_2(z, \theta_{20}, h_{20})']$ be the $d_{g_1} \times d_{g_2}$ matrix of correlations between the identifying moment function g_1 and the auxiliary moment function g_2.

From Section 3.1, let $\psi_j(z, \theta_j, h_j) = g_j(z, \theta_j, h_j) + \phi_j(z, \theta_j, h_j)$ be the locally robust moment functions corresponding to g_j $(j = 1, 2)$. Therefore, $\mathbb{E}[\psi_j(z, \theta_{j0}, h_{j0}) \psi_j(z, \theta_{j0}, h_{j0})'] = V_j$ $(j = 1, 2)$. Let $G_j^\star = \mathbb{E}[\partial \psi_j(z, \theta_{j0}, h_{j0}) / \partial \theta_j]$, $B^\star = \mathbb{E}[\psi_1(z, \theta_{10}, h_{10}) \psi_2(z, \theta_{20}, h_{20})']$ be the $d_{g_1} \times d_{g_2}$ matrix of correlations between the locally robust identifying moment function ψ_1 and the locally robust auxiliary moment function ψ_2. and $P^\star = V_2^{-1} - V_2^{-1} G_2^\star (G_2^{\star\prime} V_2^{-1} G_2^\star)^{-1} G_2^{\star\prime} V_2^{-1}$.

Theorem 4.1. *Consider the moment conditions model* (2.1), (2.2). *Under Assumptions 4.1-4.4, and given nonparametric estimators* $\hat{h}_j$ *satisfying* $||\hat{h}_j - h_{j0}|| = o_p(n^{-\frac{1}{4}})$ $(j = 1, 2)$,
(i) *the limiting distribution of the two-step GEL-weighted estimator of* θ_{10} *based on the non-locally robust moment functions is described by*

$$\sqrt{n}(\hat{\theta}_1 - \theta_{10}) \xrightarrow{d} \mathcal{N}(0, \Sigma_1 + \Sigma_2),$$

where $\Sigma_1 = G_1^{-1} V_1 G_1'^{-1}$ *and* $\Sigma_2 = G_1^{-1}(BPV_2PB' - B^\star PB' - BPB^{\star\prime}) G_1'^{-1}$**;**

(ii) *the limiting distribution of the two-step GEL-weighted estimator of* θ_{10} *based on the locally robust moment functions is described by*

$$\sqrt{n}(\hat{\theta}_1 - \theta_{10}) \xrightarrow{d} \mathcal{N}(0, \Sigma_1^\star - \Sigma_2^\star),$$

where $\Sigma_1^\star = G_1^{\star-1} V_1 G_1^{\star\prime-1}$ *and* $\Sigma_2^\star = B^\star P^\star B^{\star\prime}$.

Remark 4.1. If θ_{10} is estimated without using the auxiliary restrictions; i.e. by method of moments estimation with a plug-in $\hat{h}_1$ estimate, the asymptotic variance is given by Σ_1, which provides a basis to evaluate information contributed by auxiliary restrictions.

Remark 4.2. Theorem 4.1(i) shows that in general, two-step GEL-weighted estimation does not guarantee efficiency gains from the use of auxiliary moment restrictions with plug-in estimates. Although the use of auxiliary moment restrictions may reduce the limiting variance, Σ_2 is indefinite and depends on the various direction of correlations between moment functions and their locally robust counterparts. In the case where the adjustment terms $\phi_j(z, \theta_{j0}, h_{j0})$ $(j = 1, 2)$ are zero, it can be shown by part (ii) of the theorem that efficiency gains are guaranteed.

Remark 4.3. Theorem 4.1(ii) shows that by adjusting the auxiliary moment restriction to account for first-stage estimation error, GEL-weighted estimation guarantees efficiency gains over only using the identifying moment restrictions. The asymptotic variance is reduced by the positive definite matrix $\Sigma_2^\star$. The structure of $\Sigma_2^\star$ suggests the larger the correlation between the locally robust identifying and auxiliary moment restrictions, as captured by matrix $B^\star$, the greater the efficiency gains.

5 Semiparametric Missing Data Models

Many models in the missing data and treatment effects literature are formulated in terms of semiparametric moment conditions with the parameter of interest θ_1 often representing some causal effect. To identify θ_1 with missing data, restrictions are usually placed on the patterns of missingness. A wide class of popular models assumes that data are missing at random (MAR). MAR is a slightly less restrictive assumption than selection on observables, and is formally stated below; see Section 2.1.1, Chapter 1 of the thesis.

In this setting, in an attempt to account for missing data or treatment selection, the estimation of at least one of two nuisance functions is required. One such function is the propensity score, i.e. the probability of data missing conditional on observables. With experimental data, the propensity score may be known by design, whereas with observational data, it may need to be estimated, possibly involving conditioning on a high dimensional vector of observables to make the MAR assumption more plausible.

The other nuisance function is the expectation of the moment function conditional on observables. In the study of average treatment effects, this is simply the conditional expectation of outcomes given observed covariates, which is easy to estimate nonparametrically. However, when the moment function is some general nonlinear function of data and unknown parameters, this nuisance function is more challenging to estimate.

5.1 Moment condition models with missing data

Initially a simple, widely-used class of moment condition models with missing data is discussed. Similar set-ups are studied in Robins et al. [34], Wooldridge [40] and Graham et al. [18].

Let $z = (y, x)$ denote a vector of the observables in which y is a d_y-dimensional vector of variables that are MAR. Let d be a binary variable indicating whether or not y is observed, and x a fully observed d_x-dimensional vector of variables. The d_{θ_1}-dimensional unknown parameter of interest θ_1 uniquely satisfies the moment condition

$$\mathbb{E}[g_1(z, \theta_{10})] = 0. \tag{5.1}$$

For simplicity, θ_{10} is just-identified as in Section 4; the over-identified case gives analogous results. Let $p_0(x) = \mathbb{P}(d = 1|x)$ represent the propensity score and $q_0(x; \theta_{10}) = \mathbb{E}[g_1(z, \theta_{10})|x]$ the conditional expectation of the moment function given x.

Assumption. 5.1. **(i)** *Equation (5.1) holds uniquely at $\theta = \theta_0$;* **(ii)** *$\{z_i, d_i\}_{i=1}^n$ is an i.i.d. random sequence from the true distribution of $\{z, d\}$;* **(iii)** *the vector $\{d_i, x_i, d_i y_i\}_{i=1}^n$ only is observable;* **(iv)** *y is MAR; i.e. $d \perp y|x$;* **(v)** *for any $x \in \mathcal{X} \subset \mathbb{R}^{d_x}$, $\kappa \leq p_0(x) < 1$ for some $\kappa > 0$.*

Assumption 5.1(i) determines identification of θ_{10}. Assumption 5.1(iii) reveals what information is available to the researcher. The MAR assumption, Assumption 5.1(iv), states that missingness is not dependent on outcome values after controlling for fully observed variables x. Assumption 5.1(v) is an overlap condition common in propensity score matching, requiring neither complete observability or missingness at all values of x in the population.

Graham ([17], Theorem 2.1) shows that the total information provided by Assumption 5.1 can be summarised by the moment conditions

$$\mathbb{E}\Big[\frac{d}{p_0(x)} g_1(z, \theta_{10})\Big] = 0 \tag{5.2}$$

$$\mathbb{E}\Big[\frac{d}{p_0(x)} - 1 \,\Big|\, x\Big] = 0, \tag{5.3}$$

where the first moment restriction identifies θ_{10} by appropriately accounting for MAR data, and the second set of moment restrictions describe the propensity score. The optimal linear combination of the two sets of moment restrictions required for efficient estimation has been derived by Robins et al. [34].

In particular, an estimator $\hat{\theta}_1$ of θ_{10} that achieves the semiparametric efficiency lower bound is one that solves the moment equations

$$\frac{1}{n}\sum_{i=1}^{n} \frac{d_i g_1(z_i, \hat{\theta}_1)}{\hat{p}(x_i)} - \Big(\frac{d_i}{\hat{p}(x_i)} - 1\Big)\hat{q}(x; \hat{\theta}_1) = 0, \tag{5.4}$$

where $\hat{p}(x)$ and $\hat{q}(x, \theta_{10})$ are nonparametric estimators of $p_0(x)$ and $q_0(x, \theta_{10})$. The estimating equation (5.4) is also doubly robust. That is, given models for the propensity score $p_0(x)$ and the conditional expectation function $q_0(x, \theta_{10})$, the estimator based on (5.4) still provides a consistent estimator for θ_{10} if at most one of these two models are misspecified. In addition, if nuisance models are not specified, estimators based on doubly robust estimating equations have been shown to possess better finite sample properties (Frölich et al. [16]) and are less sensitive to choices of tuning parameters (Firpo and Rothe [15]).

5.2 Auxiliary moment restrictions

There are often small sample issues, for example, in analysis of clinical trials, that make purely nonparametric approaches undesirable. While the estimation of a conditional expectation of a single variable can be computationally straightforward, estimation of $q_0(x; \theta_{10})$ is likely to be more difficult, particularly for high dimensional x. Given the double robustness property offered by (5.4), a semiparametric approach for estimating $q_0(x; \theta_{10})$ may be a useful compromise.

5.2.1 Semiparametric missing data model

Graham [17] directly models $q_0(x;\theta_{10})$ by the *functional restriction*

$$\mathbb{E}[g_1(z,\theta_{10})|x] = q_1(x,\alpha_0,h_{10}(x_2);\theta_{10}), \tag{5.5}$$

where α_0 is a finite dimensional unknown parameter and $h_{10}(x_2)$ is an unknown function of $x_2 \subset x$. A special case includes the partial linear structure $q_1(x,\alpha_0,h_{10}(x_2);\theta_{10}) = x_1'\alpha_0 + h_{10}(x_2) - \theta_{10}$ studied in Wang et al. [39] for estimating population means under missing data. The functional restriction can be expressed in terms of the conditional moment restriction $\mathbb{E}[u(z,\alpha_0,h_{10}(x_2);\theta_{10})|x] = 0$ where $u(z,\alpha_0,h_{10}(x_2);\theta_{10}) = g_1(z,\theta_{10}) - q_1(x,\alpha_0,h_{10}(x_2);\theta_{10})$. Furthermore, due to MAR, $\mathbb{E}[u(z,\alpha_0,h_{10}(x_2);\theta_{10})|x] = \mathbb{E}[u(z,\alpha_0,h_{10}(x_2);\theta_{10})|x,d]$ and thus the conditional moment restriction may be consistently estimated using all complete case units.

For example, if q_1 is the partially linear model, then given a consistent estimator $\hat{\theta}_{HT}$ for θ_{10} obtained by the Horvitz-Thompson inverse probability weighting method, and a nonparametric estimator $\hat{h}_1(x_2)$, an estimator $\hat{\alpha}$ for α_0 solves the estimating equations

$$\frac{1}{n}\sum_{i=1}^{n} d_i a(x_i)\big(x_i'\hat{\alpha} + \hat{h}_1(x_{2i}) - g_1(z_i,\hat{\theta}_{HT})\big) = 0,$$

where $a(x)$ is a valid vector of instruments of dimension at least $\dim(\alpha)$. Details and intuition on the extent to which the restriction (5.5) provide efficiency gains is discussed in Graham ([17], pp. 446-449).

5.2.2 Auxiliary moment restriction

Here a more cautious approach is considered that guards against possible misspecification of (5.5). Following Donald and Newey [14] and Donald et el. [13], the conditional moment restriction can be estimated by a sequence of unconditional moment restrictions that increase at an appropriate rate. To facilitate analysis on behaviour under misspecification, the model considered here allows the possibility that the researcher is not able to describe $q_0(x;\theta_{10})$ perfectly, but wishes to exploit information from an auxiliary unconditional moment restriction.

For a $d_{\theta_2}-$dimensional unknown parameter $\theta_2 \in \Theta_2$ where $\Theta_2 \subset \mathbb{R}^{d_{\theta_2}}$ is compact, and unknown functions $h_2 : \mathcal{X} \times \Theta_2 \to \mathbb{R}$, consider the auxiliary moment restriction

$$\mathbb{E}[g_2(d,z,\theta_{20},h_{20})] = 0, \tag{5.6}$$

where h_{20} denotes the nuisance function evaluated at the true value θ_{20}.

Such auxiliary moment restrictions are often used in missing data models for the purposes of exploiting additional population information to improve estimation. For example, Hellerstein and Imbens [21] use census information on population averages to reduce selection bias.

Auxiliary moment restrictions may also be imposed when *surrogate data* is available, that is, when information is collected on a proxy variable closely related to a variable that has missing entries. For example, in order to study the association between obesity and high blood pressure, Qin et al. [32] model the relationship between a fully observed proxy, BMI measurements, and a more accurate measure of body fat, dual energy X-ray absorptiometry percentage, that has missing entries. For the analysis of voting behaviour in US general elections, Chen et al. [10] use post-election survey results on candidate preferences as a proxy for a non-voter's candidate choice. Any known relationships between the surrogate variable and a missing variable may be formulated as an auxiliary moment restriction.

The information that is available to the researcher is (5.2), (5.3) and (5.6). For estimation of θ_{10} using the two-step GEL-weighted approach, the following assumption is imposed.

Assumption. 5.2. *Assumption 4.1-4.4 is satisfied with ψ_1 there to $\tilde{g}_1(d, z, \theta_1, p, q) = dp(x)^{-1} \times g_1(z, \theta_1) - (d - p(x))p(x)^{-1}q(x; \theta_1)$, and ψ_2 there equal to $\tilde{g}_2(d, z, \theta_2, h_2)$ where $\tilde{g}_2$ is a locally robust version of the moment function satisfying* (5.6)*, and $q(x, \theta_1)$ is any working model for the conditional expectation function.*

$\tilde{g}_1$ is therefore the doubly robust moment function corresponding to (5.4) where a working model $q(x; \theta_{10})$ for the conditional expectation function does not need to be correct, that is, it may be the case that $q(x; \theta_{10}) \neq q_0(x; \theta_{10})$. $\tilde{g}_2$ is the locally robust version of the auxiliary moment restriction (5.6); accounting for first-step nuisance estimation of $h_{20}(x)$ allows for a guaranteed efficiency gain from exploiting the information provided in (5.6).

5.3 Two-step GEL-weighted estimation

Consider the following estimation method for θ_{10} under restrictions (5.1) and (5.6) under Assumptions 5.1 and 5.2.

Suppose $\hat{p}(x)$ and $\hat{q}(x; \theta)$ are nonparametric estimators of the unknown functions $p_0(x)$ and $q_0(x; \theta)$, respectively.

Step One: Given a nonparametric estimator $\hat{h}_2$ of h_{20}, and having obtained the locally robust moment function $\tilde{g}_2$ defined in Assumption 5.2, estimate $\tilde{g}_2$ by GEL. Collect the implied probabilities $\{\hat{\pi}_i\}_{i=1}^n$.

Step Two: Given nonparametric estimators $\hat{p}$ and $\hat{q}$ of p_0 and q_0, respectively, the two-step GEL-weighted estimator $\hat{\theta}_1$ of θ_{10} solves the estimating equation

$$\sum_{i=1}^{n} \hat{\pi}_i \tilde{g}_1(d, z, \hat{\theta}_1, \hat{p}, \hat{q}) = 0, \tag{5.7}$$

where the function $\tilde{g}_1$ is defined in Assumption 5.2.

The estimator $\hat{\theta}_1$ has the following asymptotic properties.

Corollary 5.1. *Let all nonparametric estimators involved in estimation of $\tilde{g}_1$ and $\tilde{g}_2$ converge at a rate faster than $n^{-\frac{1}{4}}$. For the moment condition model* (5.1) *and* (5.6) *under Assumptions 5.1 and 5.2,* **(i)** *$\hat{\theta}_1$ is doubly robust with respect to nuisance estimates of $p_0(x)$ and $q_0(x, \theta_{10})$ if $g_2(z, \theta_2, h_2)$ does not contain $p_0(x)$;* **(ii)** *under misspecification $q(x; \theta_{10}) \neq q_0(x; \theta_{10})$, $\hat{\theta}_1$ satisfies*

$$\sqrt{n}(\hat{\theta}_1 - \theta_{10}) \xrightarrow{p} N(0, G_1^{-1}(\Sigma_0 - \Sigma_1 - \Sigma_2)G_1'^{-1}),$$

where

$$\begin{aligned}
\Sigma_0 &= \mathbb{E}\Big[\frac{g_1(z,\theta_{10})g_1(z,\theta_{10})'}{p_0(x)}\Big] \\
\Sigma_1 &= \mathbb{E}\Big[\frac{g_1(z,\theta_{10})q(x;\theta_{10})'}{p_0(x)}\Big]\mathbb{E}\Big[\frac{g_1(z,\theta_{10})q(x;\theta_{10})'}{p_0(x)(1-p_0(x))}\Big]^{-1}\mathbb{E}\Big[\frac{g_1(z,\theta_{10})q(x;\theta_{10})'}{p_0(x)}\Big]' \\
\Sigma_2 &= \mathbb{E}\Big[\tilde{g}_2(d,z,\theta_{20},h_{20}(x))\Big(\frac{dg_1(z,\theta_{10})}{p_0(x)} - \Big(\frac{d}{p_0(x)} - 1\Big)q(x,\theta_{10})\Big)'\Big] \\
&\quad \times P_2\mathbb{E}\Big[\tilde{g}_2(d,z,\theta_{20},h_{20}(x))\Big(\frac{dg_1(z,\theta_{10})}{p_0(x)} - \Big(\frac{d}{p_0(x)} - 1\Big)q(x,\theta_{10})\Big)'\Big]'
\end{aligned}$$

where $G_1 = \mathbb{E}[\partial g_1(z, \theta_{10})/\partial\theta_1]$, $P_2 = \Omega_2^{-1} - \Omega_2^{-1}G_2(G_2'\Omega_2^{-1}G_2)^{-1}G_2'\Omega_2^{-1}$,
$\Omega_2 = \mathbb{E}[\tilde{g}_2(d, z, \theta_{20}, h_{20}(x))\tilde{g}_2(d, z, \theta_{20}, h_{20}(x))']$ and $G_2 = \mathbb{E}[\partial\tilde{g}_2(d, z, \theta_{20}, h_{20}(x))/\partial\theta_2]$.

5.4 Discussion of results

Corollary 5.1 has a clear intuitive interpretation in terms of both efficiency and robustness.

Remark 5.1. Comparisons with the Horvitz-Thompson estimator. The asymptotic variance structure shows that the proposed estimation procedure guarantees efficiency gains over Σ_0, the asymptotic variance of the original Horvitz-Thompson [23] inverse probability weighting estimator when the propensity score $p_0(x)$ is known.

Remark 5.2. Efficiency gains from propensity score estimation. The asymptotic variance is reduced by two positive definite matrices. Σ_1 represents the efficiency gain from estimating the propensity score. In particular, a nonparametric propensity score estimator incorporates information from the conditional moment restriction (5.3), which allows for efficiency gain over the Horvitz-Thompson [23] estimator. When the working model q is correctly specified so that $q(x; \theta_{10}) = q_0(x; \theta_{10})$, $\Sigma_0 - \Sigma_1$ is the asymptotic variance corresponding to the semiparametric efficiency lower bound under the missing data model in the absence of the auxiliary restriction (5.6).

Remark 5.3. Efficiency gains from auxiliary moment restrictions. The asymptotic variance is further reduced by a positive definite matrix Σ_2 representing an efficiency gain from using the auxiliary restriction (5.6). As in Theorem 4.1, the extent of the efficiency gain depends on the correlation between the moment functions $\tilde{g}_1$ and $\tilde{g}_2$, i.e. the more relevant the auxiliary moment restriction is, the lower the asymptotic variance.

Remark 5.4. Relative efficiency under misspecification. The semiparametric efficiency lower bound for the moment condition model (5.2), (5.3) and (5.5) is derived by Graham [17]. However, when the functional restriction (5.5) is misspecified, that is when $\mathbb{E}[g_1(z,\theta_{10})|x] \neq q_1(x,\alpha_0,h_{10}(x_2);\theta_{10})$, the estimator of θ_{10} based on the efficient influence function leads to an asymptotic variance no lower than $\Sigma_0 - \Sigma_1$, see p.449, discussion of Proposition 3.2 of Graham [17]. Here, the asymptotic variance is reduced further by Σ_2, suggesting relative efficiency under misspecification. However, for the moment condition model (5.2), (5.3) and (5.6), the asymptotic variance of the estimator of θ_{10} based on (5.7) will not, in general, coincide with the semiparametric efficiency lower bound.

Remark 5.5. Double robustness I. The usual double robustness properties in this missing data set-up are preserved here. In particular, if either the model for the propensity score $p_0(x)$ or the conditional expectation function $q_0(x,\theta_{10})$ is misspecified, the estimator of θ_{10} remains consistent. However, the auxiliary moment restriction (5.6) is required to be correctly specified. In practice, as in the case of surrogate variables discussed in Section 5.2.2, the auxiliary restriction is likely to model a conditional relationship between z and x and so will remain valid under the MAR assumption. In contrast, if estimation of θ_{10} is based on Graham [17]'s efficient influence function under (5.2), (5.3) and (5.5), more stringent requirements for double robustness are needed. In particular, those variables entering $q_1(x,\alpha_0,h_0(x_2);\theta_{10})$ parametrically should not affect either the propensity score $p_0(x)$ or the conditional variance of the moment function g_1, $\mathbb{E}[g_1(z,\theta_{10})g_1(z,\theta_{10})'|x]$, see p.449 of Graham [17]. Since propensity score estimation often involves conditioning on all covariate information in order to make the MAR assumption plausible, the double robustness result in Graham [17] may not hold in such instances.

Remark 5.6. Double robustness II. Double robustness properties are still valuable even if nuisance functions are estimated nonparametrically. Since doubly robust moment functions retain mean-zeroness for movements of nuisance estimates around the true value, Monte Carlo evidence in Frölich et al. [16] shows that estimators based on doubly robust estimating equations are less sensitive to choices of tuning parameters. Furthermore, Firpo and Rothe [15] show that for polynomial kernel estimation of the propensity score, larger smoothing biases are permitted while maintaining the faster than $n^{-\frac{1}{4}}$ convergence rate required.

Remark 5.7. Efficiency-robustness trade-off. These comparisons highlight potential trade-offs that exist between efficiency gains from using auxiliary moment restrictions versus robustness properties. While for some treatment effects and missing data models under MAR the efficient influence function is doubly robust, an orthogonality condition with respect to individual nuisance functions is not guaranteed for more complicated models that may be tailored for specific applications.

6 Simulation study: using surrogate information on missing outcomes

In this section, we consider the problem of estimating the expectation of a random variable that is MAR. This is a similar set-up to that studied in Section 4, Chapter 1 of the thesis, however, we investigate the finite sample performance of the GEL-weighted estimator studied in this paper. Moreover, our main focus is to examine the double robustness property of the GEL-weighted estimator under misspecification of the conditional expectation function or propensity score. Overall, the simulation study shows the performance of the GEL-weighted estimator is promising in small samples and often displays advantages in terms of mean square error over popular alternatives suggested in the literature.

Let $x \sim N(1,1)$ be a covariate, and $y = 0.25x + \epsilon$ be an outcome variable, where ϵ is an error term. Let d be a binary variable such that $d = 1$ if and only if y is observed. The propensity score is given by $\mathbb{P}(d = 1|x) = \exp(-1 + x)/(1 + \exp(-1 + x))$.

In addition, data on a surrogate variable is available. Let $s = 0.5x + u$, where u is an error term. The error terms ϵ and u are jointly normal such that

$$\begin{pmatrix} \epsilon \\ u \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right).$$

Therefore, the further ρ is from zero, the more correlated s is with y. The auxiliary moment restrictions that describe the mean and symmetry of the distribution of s are given by

$$\mathbb{E}[g_2(s, \theta_{20})] = \mathbb{E}\begin{bmatrix} s - \theta_{20} \\ (s - \theta_{20})^3 \end{bmatrix} = 0.$$

$\theta_{20} = \mathbb{E}[s] = 0.5$ is assumed to be unknown to the researcher and must be estimated. The parameter of interest is $\theta_{10} = \mathbb{E}[y] = 0.25$. The following estimators of θ_{10} are considered for comparison.

6.1 Estimators of θ_{10}

Let $\hat{p}(x)$ be the estimated propensity score from a logistic regression of d on $(1, x)$, and $\hat{q}(x)$ be the fitted values from a regression of y on x using the set of observations for which $d_i = 1$, $(i = 1, ..., n)$.

The IPW estimator of θ_{10} is given by $\hat{\theta}_{1,IPW} = n^{-1} \sum_{i=1}^{n} \hat{p}(x_i)^{-1} d_i y_i$, and the doubly robust estimator (DR) is given by $\hat{\theta}_{1,DR} = (n^{-1} \sum_{i=1}^{n} d_i \hat{p}(x_i)^{-1} y_i) - (n^{-1} \sum_{i=1}^{n} \hat{p}(x_i)^{-1} (d_i - \hat{p}(x_i))\hat{q}(x_i))$. Both of these estimators do not use the auxiliary information given by the restriction $\mathbb{E}[g_2(s, \theta_{20})] = 0$.

The type of GEL estimator we employ in our simulation study is the continuously-updating GMM estimator (CUE). The CUE-weighted estimator (CUEW) of θ_{10} is given by

$$\hat{\theta}_{1,CUEW} = \sum_{i=1}^{n} \hat{\pi}_i \Big(\frac{d_i y_i}{\hat{p}(x_i)} - \Big(\frac{d_i}{\hat{p}(x_i)} - 1 \Big) \hat{q}(x_i) \Big),$$

where $\hat{\pi}_i$ are the CUE implied probabilities from CUE estimation of the moment condition $\mathbb{E}[g_2(s, \theta_{20})] = 0$.

Finally, for comparison we also consider a CUE estimator which jointly estimates all the available information. The joint-CUE estimator (JCUE) of θ_{10} that is based on the stacked moment conditions

$$\mathbb{E} \begin{bmatrix} \frac{dy}{\hat{p}(x)} - \Big(\frac{d}{\hat{p}(x)} - 1 \Big) \hat{q}(x) - \theta_{10} \\ s - \theta_{20} \\ (s - \theta_{20})^3 \end{bmatrix} = 0$$

is denoted $\hat{\theta}_{1,JCUE}$.

6.2 Quantifying the relevance of the auxiliary moment restriction

Since larger values of $|\rho|$ imply that s and y are more correlated, it is expected that exploiting any information on s will lead to more efficient estimation of $\mathbb{E}[y]$ when ρ is further away from zero. By the asymptotic variance results presented in Sections 4 and 5, we can quantify the efficiency gain from using auxiliary moment restriction $\mathbb{E}[g_2(s, \theta_{20})] = 0$ for the CUEW estimator.

In particular, by Theorem 4.1(ii) (also see Remarks 4.3 and 5.3) the extent of the efficiency gain from using the auxiliary moment restriction is dependent on the covariance between the identifying and auxiliary moment restrictions. The larger the covariance, the greater the efficiency gain. Let $\hat{\theta}_{20}$ be the CUE estimator based on the moment condition $\mathbb{E}[g_2(s, \theta_{20})] = 0$. A consistent estimator for this 2×1 covariance matrix B is given by

$$\hat{B}^\star = \frac{1}{n} \sum_{i=1}^{n} g_2(s_i, \hat{\theta}_{20}) \Big(\frac{d_i y_i}{\hat{p}(x_i)} - \Big(\frac{d_i}{\hat{p}(x_i)} - 1 \Big) \hat{q}(x_i) - \hat{\theta}_{1,CUEW} \Big),$$

and a consistent estimator of the efficiency gain is given by $\hat{B}^{\star\prime} \hat{P}_2 \hat{B}^\star$, where $\hat{P}_2$ is a consistent estimator for P_2 obtained by combining sample Jacobian and covariance matrices evaluated at $\hat{\theta}_{20}$ (see Corollary 5.1).

Figure 1 illustrates how the value of $\hat{B}^{\star\prime} \hat{P}_2 \hat{B}^\star$, which represents the estimated variance reduction from exploiting auxiliary information, varies with the correlation parameter ρ. The standard error estimates, based on the asymptotic results in Sections 4 and 5, of CUEW and DR are displayed for sample size $n = 2,000$, averaged over $5,000$ simulations. The large sam-

ple size is used to display asymptotic behaviour. The values of ρ are selected from $\{0.1, ..., 0.9\}$ in 0.1 intervals.

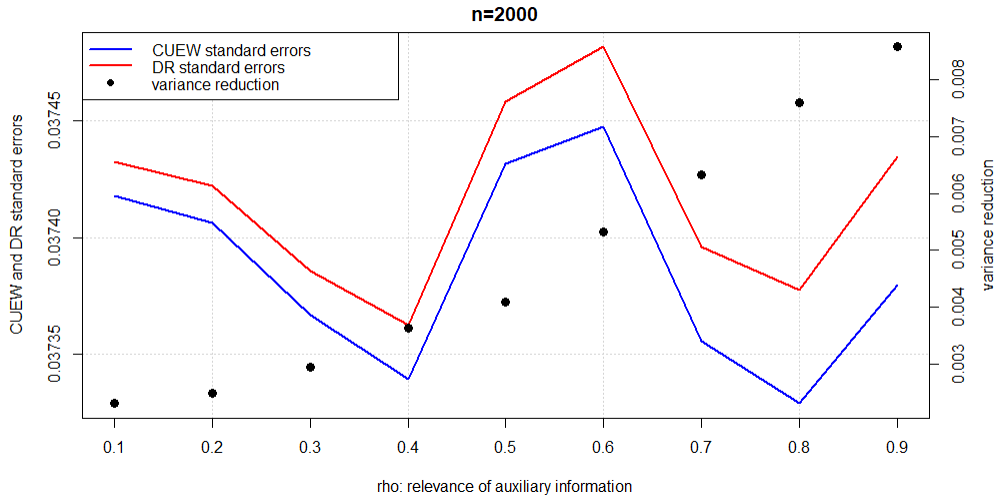


Figure 1. Relevance of auxiliary information. DR and CUEW standard errors as ρ varies. Variance reduction estimates are given by $\hat{B}^{\star\prime}\hat{P}_2\hat{B}^{\star}$.

As the correlation between the surrogate variable s and the outcome variable y increases, the estimated variance reduction increases. The standard errors of CUEW are lower than DR, and the difference between the standard errors increases as ρ increases. This displays the efficiency advantage that CUEW has over DR by using the auxiliary moment restriction, and how this advantage is greater for higher values of ρ.

6.3 Small sample performance of CUEW under correct specification

The simulation experiment notes the bias and MSE of the estimators discussed above under sample sizes $n = 100$, 200 and 400. The graphs below show the averaged results over 10,000 experiments.

Figure 2 shows that the class of inverse probability weighted estimators considered here generally display lower biases as the sample size increases. Under correct model specification, there may not be advantages in terms of finite sample bias properties from exploiting auxiliary information when nuisance parameters have to be estimated. Even for $n = 400$, estimators that do not use auxiliary information perform better in terms of bias than CUEW and JCUE even when s is highly correlation with y ($\rho = 0.7, 0.8$).

JCUE appears to be considerably more biased than the other estimators in small samples. JCUE involves jointly estimating two parameters θ_{10} and θ_{20} in an over-identified system of equations which may introduce greater biases in small samples. Importantly, CUEW only involves estimating θ_{20} to construct the implied probability weights which may explain its more

stable performance in small samples. However, the differences in bias between the estimators clearly start to diminish as expected for the higher sample size $n = 400$.

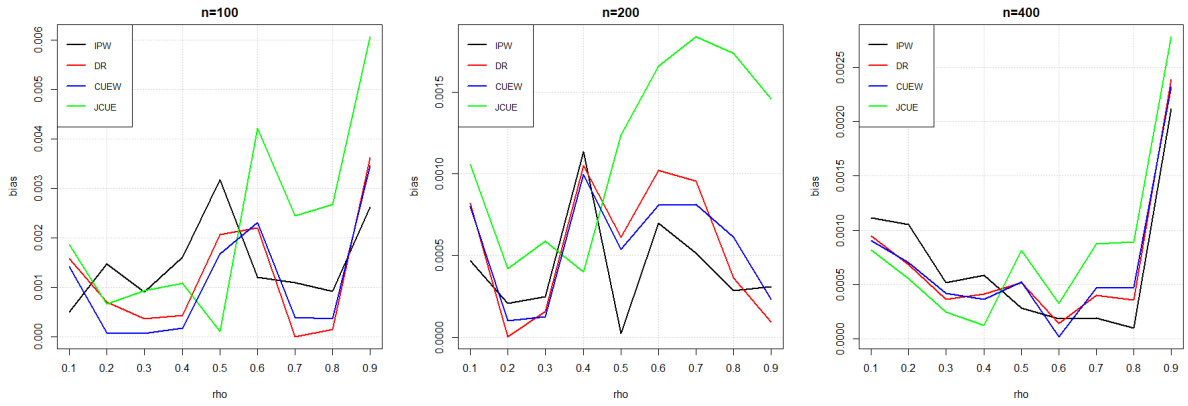


Figure 2. Bias under correct specification as ρ varies for $n = 100$, 200 and 400.

Considering the very competitive performance of IPW in terms of small sample bias, the poorer MSE properties of IPW are due to the larger variance of IPW compared with DR, CUEW and JCUE. In terms of asymptotic variance, DR is more efficient than IPW, and CUEW and JCUE are more efficient than DR. That said, Figure 3 shows that for the smallest sample $n = 100$, DR may have the best MSE properties at certain values of ρ, despite not incorporating information from the auxiliary moment restriction.

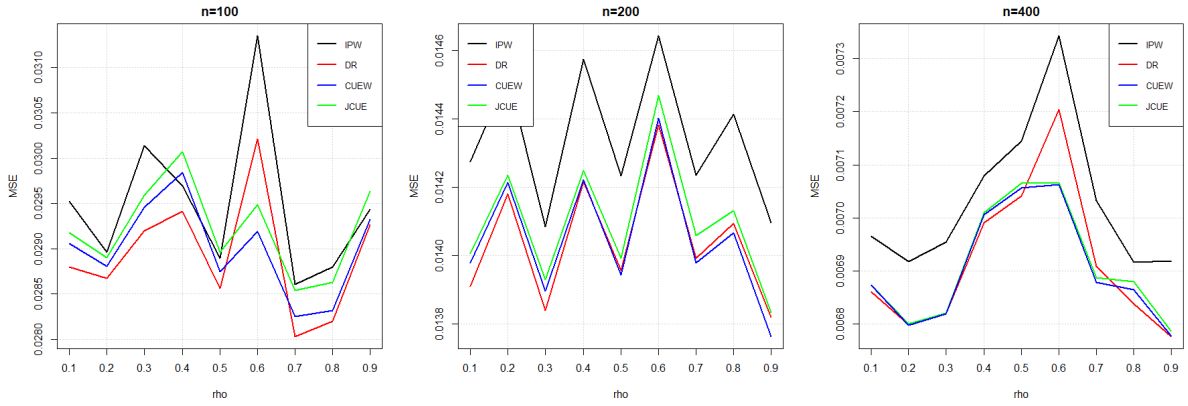


Figure 3. MSE under correct specification as ρ varies for $n = 100$, 200 and 400.

For larger sample sizes, when the auxiliary information is highly relevant ($\rho > 0.7$), CUEW and JCUE generally perform well, which given the better bias properties of DR, reflects their relative asymptotic efficiency advantages. Interestingly, compared with JCUE, the performance of CUEW is promising given that it performs better in terms of MSE and has considerably lower and stable small sample biases. This mirrors the favourable finite-sample results of GEL-weighted estimators presented in Bravo [6]; in our case, however, the auxiliary moment restriction involves nuisance estimation.

6.4 Misspecification of the conditional expectation function (CIF)

Here we consider the performance of CUEW under misspecification of the conditional expectation function. Instead of the model set-up introduced at the start of the section, now let $y = 0.25[(1-\kappa)x + \kappa(x - \mathbb{E}[x])^2] + \epsilon$ for some $0 < \kappa < 1$, where x and ϵ are as defined before. Therefore, the true value of the parameter of interest has not changed since

$$\begin{aligned}
\theta_{10} &= \mathbb{E}[y] \\
&= 0.25\big[(1-\kappa)\mathbb{E}[x] + \kappa\mathbb{E}[(x - \mathbb{E}[x])^2]\big] + \mathbb{E}[\epsilon] \\
&= 0.25[(1-\kappa) + \kappa] + 0 \\
&= 0.25.
\end{aligned}$$

However, the CIF is quadratic in x; $\mathbb{E}[y|x] = 0.25(1-\kappa)x + 0.25\kappa(x - \mathbb{E}[x])^2$. Therefore, a linear model specified for $\mathbb{E}[y|x]$ is incorrect and a larger κ represents a higher degree of misspecification. Due to the double robustness property, DR, CUEW and JCUE are consistent even if the CIF is misspecified. We now present small sample bias and MSE properties under varying degree of CIF model misspecification. The values of κ are taken from the set $\{0.1, ..., 0.9\}$ in 0.1 intervals.

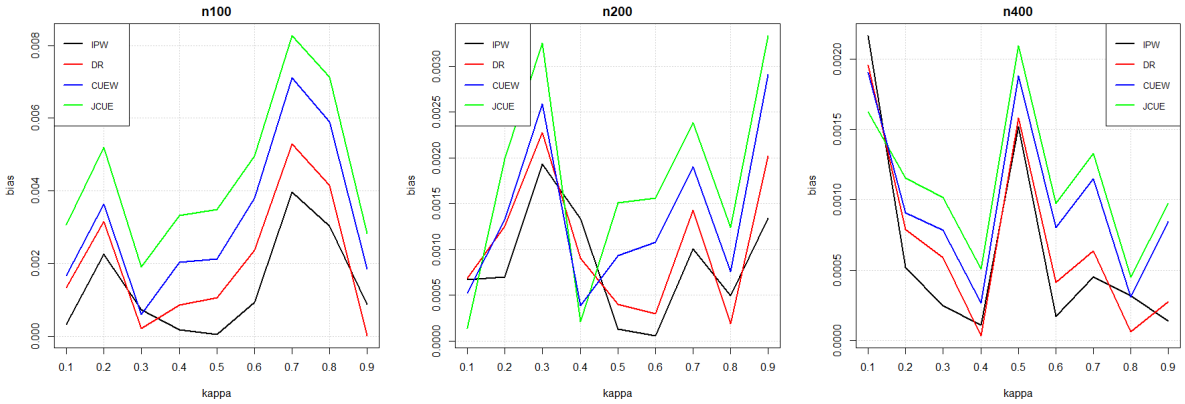


Figure 4. Bias under CIF misspecification as κ varies for $n = 100$, 200 and 400.

Figure 4 shows that the biases of DR, CUEW and JCUE generally decrease as the sample size increases, which illustrates the double robustness property that suggests all estimators should be consistent. IPW is not a function of the estimated CIF and hence is unaffected by misspecification of the CIF; it can be seen that the finite sample biases are generally lower for IPW compared with other estimators which rely on estimates of the CIF. Therefore, the idea that DR, CUEW and JCUE are negatively impacted by misspecification of the CIF relative to IPW can be seen by IPW displaying considerably lower biases than DR, CUEW and JCUE for higher values of κ; only DR is competitive with IPW in this regard.

DR is generally less biased than CUEW and JCUE, with this gap also widening as the degree of misspecification increases. CUEW is generally less biased than JCUE; as discussed in

Section 3.4 of Chapter 1, pp.26-27, misspecification of nuistance models may result in unstable estimation for one-step approaches based joint estimation of all available moment restrictions.

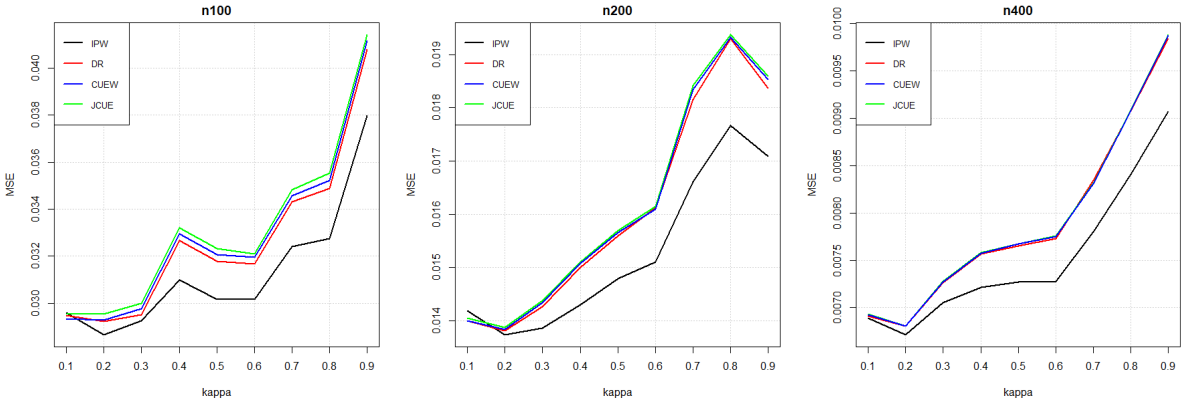


Figure 5. MSE under CIF misspecification as κ varies for $n = 100$, 200 and 400.

For all estimators, for a fixed sample size, MSE properties worsen as the degree of misspecification κ increases, but IPW performs the best in this regard. Even though IPW does not depend on estimates of the CIF, y depends heavily on the variance of x as CIF misspecification increases; these higher variances are reflected with larger MSEs reported as a function of κ.

However, in general, the MSE of all estimators decrease with the sample size for any degree of CIF misspecification. The similar MSE results of DR, CUEW and JCUE highlights the advantage that CUEW and JCUE have by using the auxiliary moment restriction over DR in terms of lower variance, which counters the relatively larger biases for CUEW and JCUE displayed in Figure 4. MSE is generally marginally lower for CUEW compared with JCUE for all sample sizes and values of κ considered here which again illustrates the potential robustness properties of two-step estimation.

6.5 Misspecification of the propensity score model

Here, instead of the model set-up introduced at the start of the section, the true propensity score model is given by

$$\mathbb{P}(d = 1|x) = \frac{\exp(-1 + x + \kappa x^2)}{1 + \exp(-1 + x + \kappa x^2)}, \ \kappa > 0,$$

whereas the estimated propensity score fits a logit regression assuming $\kappa = 0$. Thus, higher values of κ represent a greater degree of model misspecification. We now present small sample bias and MSE properties for varying degrees of misspecification of the propensity score model. The values of κ are taken from the set $\{0.05, ..., 0.45\}$ in 0.05 intervals.

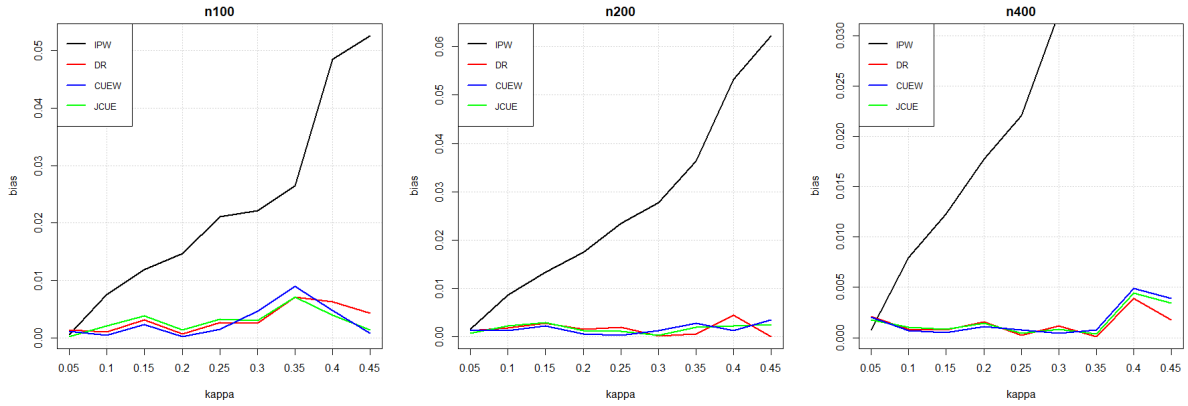


Figure 6. Bias under propensity score misspecification as κ varies for $n = 100$, 200 and 400.

While the biases of DR, CUEW and JCUE decrease with the sample size, the bias of IPW increases with the sample size. Due to the double robustness property, DR, CUEW and JCUE are consistent under this setting, however, IPW is inconsistent.

CUEW is competitive with DR and JCUE in terms of finite-sample bias properties under this setting, with no estimator exhibiting uniformly (in κ) lower biases. Although, for the highest sample size $n = 400$, DR is significantly less biased for higher levels of misspecification.

Figure 7 shows that, as for the case of CIF misspecification, the MSE of estimators generally worsens as the degree of propensity score misspecification κ increases. However, for DR, CUEW and JCUE, the MSE is generally decreasing with the sample size.

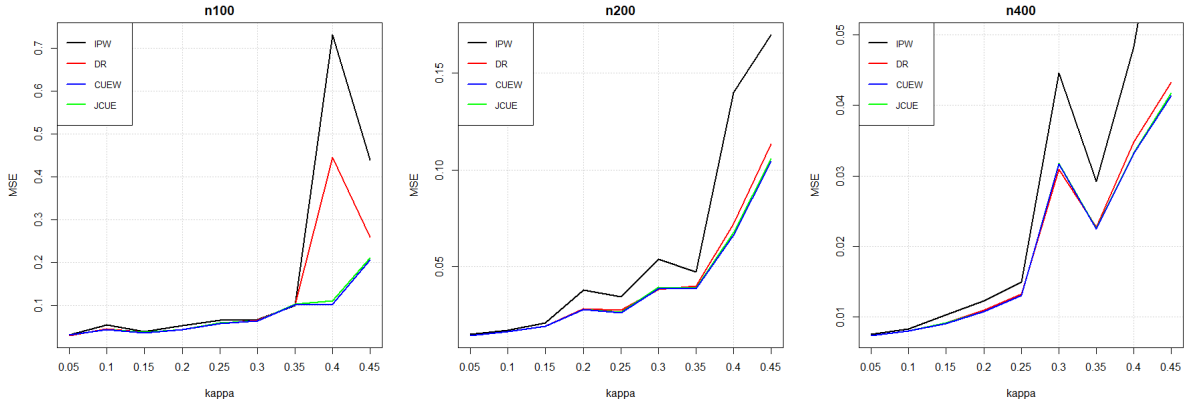


Figure 7. MSE under propensity score misspecification as κ varies for for $n = 100$, 200 and 400.

CUEW appears to display the best MSE properties over all sample sizes and degree of misspecification, however this advantage is very small, with DR, CUEW and JCUE possessing very similar MSE rates for larger sample sizes. In general, CUEW and JCUE perform better in terms of MSE than DR, suggesting it may be especially beneficial to utilise the auxiliary information when the propensity score model is misspecified.

7 Conclusion

This paper considers two-step GEL-weighted estimation designed to exploit information from auxiliary moment restrictions to guarantee an efficiency gain over identifying moment restrictions. Auxiliary moment restrictions may involve nuisance estimation. In the case where plug-in nonparametric estimates are involved, a local robustness correction (Chernozhukov et al. [12]) is required for first-stage estimation of the auxiliary moment restriction.

The results are applied to a semiparametric missing data model to show that the two-step GEL-weighted estimator preserves a double robustness property and may allow efficiency gains under misspecification of either the propensity score or conditional expectation function as compared with estimators based on the efficient influence function derived in Graham [17]. For the case where the auxiliary moment restriction consists of some population information related to a variable that is MAR, simulation results show the GEL-weighted approach has merits in terms of MSE properties in small samples, especially when the propensity score model is misspecified.

The results of this paper may be extended and generalised to further investigate the usefulness of separately estimating and combining information from identifying and auxiliary semiparametric moment restrictions. For example, locally robust corrections may not be required if plug-in estimators are not used. If sieve-GEL methods (Otsu [29]) that jointly estimate finite-dimensional parameters and unknown functions are used for estimation of the auxiliary restriction, a result similar to Theorem 4.1(ii) may hold without locally robust corrections.

Finally, since auxiliary information, such as the restrictions defining the propensity score, is often more precisely summarised by a conditional moment restriction, it would also be of interest to extend the results of this paper to the conditional moment restrictions setting. GEL implied probabilities from estimation of conditional moment restrictions can be obtained in two ways, depending on whether local GEL methods (Kitamura et al. [24] and Smith [38]) or series approximation methods (Otsu [29] and Parente and Smith [30]) are used. However, a method by which to efficiently combine information from an auxiliary conditional moment restriction with an identifying unconditional moment restriction via GEL implied probabilities is not immediately clear.

References

[1] Daniel Ackerberg, Xiaohong Chen, Jinyong Hahn, and Zhipeng Liao. Asymptotic efficiency of semiparametric two-step GMM. *Review of Economic Studies*, 2014.

[2] Chunrong Ai and Xiaohong Chen. Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions. *Econometrica*, 71(6):1795–1843, 2003.

[3] Donald W K Andrews. Asymptotics for Semiparametric Econometric Models Via Stochastic Equicontinuity. *Econometrica*, 62(1):43–72, 1994.

[4] Bertille Antoine, Hélène Bonnal, and Eric Renault. On the efficient use of the informational content of estimating equations: Implied probabilities and Euclidean empirical likelihood. *Journal of Econometrics*, 138(2):461–487, 2007.

[5] Kerry Back and David P. Brown. Implied Probabilities in GMM Estimators. *Econometrica*, 61(4):971–975, 1993.

[6] Francesco Bravo. Efficient M-estimators with auxiliary information. *Journal of Statistical Planning and Inference*, 140(11):3326–3342, 2010.

[7] Francesco Bravo, Juan Carlos Escanciano, and Ingrid Van Keilegom. Wilks' Phenomenon in Two-Step Semiparametric Empirical Likelihood Inference. 16, 2015.

[8] Bryan W. Brown and Whitney K. Newey. Efficient Semiparametric Estimation of Expectations. *Econometrica*, 66(2):453–464, 1998.

[9] Bryan W. Brown and Whitney K. Newey. Generalized method of moments, efficient bootstrapping, and improved inference. *Journal of Business and Economic Statistics*, 20(4):507–517, 2002.

[10] Song Xi Chen, Denis H Y Leung, and Jing Qin. Improving semiparametric estimation by using surrogate data. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 70(4):803–823, 2008.

[11] Xiaohong Chen, Oliver Linton, and Ingrid Van Keilegom. Estimation of semiparametric models when the criterion function is not smooth. *Econometrica*, 71(5):1591–1608, 2003.

[12] Victor Chernozhukov, Juan Carlos Escanciano, Hidehiko Ichimura, and Whitney K. Newey. Locally Robust Semiparametric Estimation. 2016.

[13] Stephen G. Donald, Guido W. Imbens, and Whitney K. Newey. Empirical likelihood estimation and consistent tests with conditional moment restrictions. *Journal of Econometrics*, 117(1):55–93, 2003.

[14] Stephen G. Donald and Whitney K. Newey. Choosing the Number of Instruments. *Econometrica*, 69(5):1161–1191, 2001.

[15] Sergio Firpo and Christoph Rothe. Properties of Doubly Robust Estimators when Nuisance Functions are Estimated Nonparametrically. 2016.

[16] Markus Frölich, Martin Huber, and Manuel Wiesenfarth. The finite sample performance of semi- and non-parametric estimators for treatment effects and policy evaluation. *Computational Statistics and Data Analysis*, 115:91–102, 2017.

[17] Bryan S. Graham. Efficiency Bounds for Missing Data Models With Semiparametric Restrictions. *Econometrica*, 79(2):437–452, 2011.

[18] Bryan S. Graham, Cristine Campos De Xavier Pinto, and Daniel Egel. Inverse probability tilting for moment condition models with missing data. *Review of Economic Studies*, 79(3):1053–1079, 2012.

[19] Jinyong Hahn. On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects. *Econometrica*, 66(2):315–331, 1998.

[20] Lars Peter Hansen, John Heaton, and Amir Yaron. Finite-sample properties of some alternative GMM estimators. *Journal of Business & Economic Statistics*, 14(3):262–280, 1996.

[21] Judith K. Hellerstein and Guido W. Imbens. Imposing Moment Restrictions from Auxiliary Data by Weighting. *The Review of Economics and Statistics*, 81(1):1–14, 1999.

[22] Nils Lid Hjort, Ian W. McKeague, and Ingrid Van Keilegom. Extending the scope of empirical likelihood. *Annals of Statistics*, 37(3):1079–1111, 2009.

[23] D. G. Horvitz and D. J. Thompson. A Generalization of Sampling Without Replacement from a Finite Universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

[24] Yuichi Kitamura, Gautam Tripathi, and Hyungtaik Ahn. Empirical likelihood-based inference in conditional moment restriction models. *Econometrica*, 72(6):1667–1714, 2004.

[25] Whitney K. Newey. The Asymptotic Variance of Semiparametric Estimators. *Econometrica*, 62(6):1349–1382, 1994.

[26] Whitney K. Newey and Daniel McFadden. Large sample estimation and hypothesis testing. *Handbook of Econometrics*, 4:2111–2245, 1994.

[27] Whitney K. Newey, Joaquim J S Ramalho, and Richard J. Smith. Identification and Inference for Econometric Models Asymptotic Bias for GMM and GEL Estimators with Estimated Nuisance Parameters. In *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, pages 245–281. 2005.

[28] Whitney K. Newey and Richard J. Smith. Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica*, 72(1):219–255, 2004.

[29] Taisuke Otsu. Empirical Likelihood Estimation of Conditional Moment Restriction Models With Unknown Functions. *Econometric Theory*, 27(01):8–46, 2011.

[30] Paulo M.D.C. Parente and Richard J. Smith. Tests of additional conditional moment restrictions. *Journal of Econometrics*, 200(1):1–16, 2017.

[31] Artem Prokhorov and Peter Schmidt. GMM redundancy results for general missing data problems. *Journal of Econometrics*, 151(1):47–55, 2009.

[32] Jing Qin, Biao Zhang, and Denis H. Y. Leung. Empirical Likelihood in Missing Data Problems. *Journal of the American Statistical Association*, 104(488):1492–1503, 2009.

[33] Joaquim J S Ramalho and Richard J. Smith. Goodness of Fit Tests for Moment Condition Models. 2006.

[34] James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of Regression Coefficients When Some Regressors Are Not Always Observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994.

[35] Daniel O. Scharfstein, Andrea Rotnitzky, and James M. Robins. Adjusting for Nonignorable Drop-Out Using Semiparametric Nonresponse Models. *Journal of the American Statistical Association*, 94(448):1096–1120, 1999.

[36] Susanne M. Schennach. Point estimation with exponentially tilted empirical likelihood. *Annals of Statistics*, 35(2):634–672, 2007.

[37] Richard J. Smith. Alternative Semi-Parametric Likelihood Approaches to Generalised Method of Moments Estimation. *The Economic Journal*, 107(441):503–519, 1997.

[38] Richard J. Smith. Efficient information theoretic inference for conditional moment restrictions. *Journal of Econometrics*, 138(2):430–460, 2007.

[39] Qihua Wang, Oliver Linton, and Wolfgang Härdle. Semiparametric Regression Analysis With Missing Response at Random. *Journal of the American Statistical Association*, 99(466):334–345, 2004.

[40] Jeffrey M. Wooldridge. Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics*, 141(2):1281–1301, 2007.

Appendix

For $j = 1, 2$, and $i = 1, ..., n$: $g_{ji}(\theta_j, h_j) = g_j(z_i, \theta_j, h_j)$; $g_{ji} = g_{ji}(\theta_{j0}, h_{j0})$; $\hat{g}^h_{ji} = g_{ji}(\theta_{j0}, \hat{h}_j)$; $\hat{g}_{ji} = g_{ji}(\hat{\theta}_j, \hat{h}_j)$; $\hat{g}_j(\theta, h) = \sum_{i=1}^n g_{ji}(\theta, h)/n$; $\hat{g}_j = \sum_{i=1}^n \hat{g}_{ji}/n$; $\hat{g}_j(\theta_j, h_j) = \sum_{i=1}^n g_{ji}(\theta_j, h_j)/n$; $\hat{g}^h_{ji} = g_{ji}(\theta_{j0}, \hat{h}_j)$; $G_{ji}(\theta_j, h_j) = G_j(z_i, \theta_j, h_j)$; $G_{ji} = G_{ji}(\theta_{j0}, h_{j0})$; $\hat{G}^h_{ji} = G_{ji}(\theta_{j0}, \hat{h}_j)$; $\hat{G}_{ji} = G_{ji}(\hat{\theta}_j, \hat{h}_j)$.

A Proof of main results

Proof of Theorem 4.1(i): Consistency

Rewrite

$$\sum_{i=1}^n \hat{\pi}_i \hat{g}_{1i} = -\frac{1}{n}\sum_{i=1}^n \hat{\rho}_{2i}\hat{g}_{1i} + R_n,$$

where $\hat{\rho}_{2i} = \rho_2(\hat{\lambda}'\hat{g}_{2i})$ $(i = 1, ..., n)$ and

$$\begin{aligned} R_n &= \Big(\frac{n}{\sum_{j=1}^n \hat{\rho}_{2j}} + 1\Big)\frac{1}{n}\sum_{i=1}^n \hat{\rho}_{2i}\hat{g}_{1i} \\ &= \Big(\frac{n}{\sum_{j=1}^n \hat{\rho}_{2j}} + 1\Big)\frac{1}{n}\sum_{i=1}^n (\hat{\rho}_{2i} + 1)\hat{g}_{1i} - \Big(\frac{n}{\sum_{j=1}^n \hat{\rho}_{2j}} + 1\Big)\frac{1}{n}\sum_{i=1}^n \hat{g}_{1i}. \end{aligned}$$

By Lemma B.1, $\hat{\rho}_{2i} + 1 = o_p(1)$ uniformly $(i = 1, ..., n)$. Therefore, $(n/\sum_{j=1}^n \hat{\rho}_{2j}) + 1 = o_p(1)$.

Hence,

$$\begin{aligned} ||R_n|| &\leq \Big|\Big|\frac{n}{\sum_{j=1}^n \hat{\rho}_{2j}} + 1\Big|\Big|\Big(\big[\max_{1\leq i\leq n} |\hat{\rho}_{2i} + 1| + 1\big]\frac{1}{n}\sum_{i=1}^n ||\hat{g}_{1i}||\Big) \\ &\leq \Big|\Big|\frac{n}{\sum_{j=1}^n \hat{\rho}_{2j}} + 1\Big|\Big|\big[\max_{1\leq i\leq n} |\hat{\rho}_{2i} + 1| + 1\big]\Big(\frac{1}{n}\sum_{i=1}^n d_1(z_i)\Big) \\ &= o_p(1)\big[1 + o_p(1)\big]\big(\mathbb{E}[d_1(z)] + o_p(1)\big) \\ &= o_p(1), \end{aligned}$$

where the first inequality follows by CS and T, the second inequality follows by Assumption 4.3(i), the first equality follows by the above arguments, WLLN and Lemma B.1, and the second equality follows by Assumption 4.3(i).

Hence,

$$\Big|\Big|\sum_{i=1}^n \hat{\pi}_i \hat{g}_{1i} + \frac{1}{n}\sum_{i=1}^n \hat{\rho}_{2i}\hat{g}_{1i}\Big|\Big| = o_p(1). \tag{A.1}$$

Now,

$$\begin{aligned}\Big|\Big| -\frac{1}{n}\sum_{i=1}^{n}\hat{\rho}_{2i}\hat{g}_{1i} - \frac{1}{n}\sum_{i=1}^{n}\hat{g}_{1i}\Big|\Big| &= \Big|\Big| -\frac{1}{n}\sum_{i=1}^{n}(\hat{\rho}_{2i}+1)\hat{g}_{1i}\Big|\Big| \\ &\leq \max_{1\leq i\leq n}|\hat{\rho}_{2i}+1|\Big(\frac{1}{n}\sum_{i=1}^{n}d_1(z_i)\Big) \\ &= o_p(1)\big(\mathbb{E}[d_1(z)]+o_p(1)\big) \\ &= o_p(1), \end{aligned} \tag{A.2}$$

where the inequality follows by CS and Assumption 4.3(i), the second equality follows by Lemma B.1 and WLLN, and the third inequality follows by Assumption 4.3(i).

By T, $(A.1)$, $(A.2)$, and since $\hat{\theta}_1$ solves $\sum_{i=1}^{n}\hat{\pi}_i\hat{g}_{1i}=0$,

$$\Big|\Big|\frac{1}{n}\sum_{i=1}^{n}\hat{g}_{1i}\Big|\Big| \leq o_p(1). \tag{A.3}$$

For any $\theta_1\in\Theta_1$,

$$\begin{aligned}\Big|\Big|\frac{1}{n}\sum_{i=1}^{n}(g_{1i}(\theta_1,\hat{h}_1)-g_{1i}(\theta_1,h_{10}))\Big|\Big| &\leq \Big(\frac{1}{n}\sum_{i=1}^{n}b_1(z_i)\Big)||\hat{h}_1-h_{10}|| \\ &= \big(\mathbb{E}[b_1(z)]+o_p(1)\big)||\hat{h}_1-h_{10}|| \\ &\leq o_p(1), \end{aligned}$$

where the first inequality follows by CS and Assumption 4.3(iii), the equality by WLLN, and the second inequality follows by Assumption 4.3(iii).

Thus, by T and $(A.3)$,

$$\Big|\Big|\frac{1}{n}\sum_{i=1}^{n}g_{1i}(\hat{\theta}_1,h_{10})\Big|\Big| \leq o_p(1).$$

For $\theta_1\in\Theta_1$, let $Q_n(\theta_1)=||\sum_{i=1}^{n}g_{1i}(\theta_1,h_{10})/n||$ and $Q_0(\theta_1)=||\mathbb{E}[g_1(\theta_1,h_{10})]||$. Therefore, $Q_n(\hat{\theta}_1)\leq o_p(1)$. Also, by Assumption 4.1(i), $Q_0(\theta_{10})=0$, and $\kappa=\inf_{\theta_1\in\Theta_1,||\theta_1-\theta_{10}||>\epsilon}Q_0(\theta_1)>0$. Then,

$$\mathbb{P}\big(||\hat{\theta}_1-\theta_{10}||>\epsilon\big)\leq\mathbb{P}\big(Q_0(\hat{\theta}_1)\geq\kappa\big). \tag{A.4}$$

Write

$$\begin{aligned}\mathbb{P}(Q_0(\hat{\theta}_1)\geq\kappa) &= \mathbb{P}\Big(Q_0(\hat{\theta}_1)\geq\kappa\Big|\sup_{\theta_1\in\Theta_1}|Q_n(\theta_1)-Q_0(\theta_1)|\leq\frac{\kappa}{2}\Big)\times\mathbb{P}\Big(\sup_{\theta_1\in\Theta_1}|Q_n(\theta_1)-Q_0(\theta_1)|\leq\frac{\kappa}{2}\Big) \\ &\quad +\mathbb{P}\Big(Q_0(\hat{\theta}_1)\geq\kappa\Big|\sup_{\theta_1\in\Theta_1}|Q_n(\theta_1)-Q_0(\theta_1)|>\frac{\kappa}{2}\Big)\times\mathbb{P}\Big(\sup_{\theta_1\in\Theta_1}|Q_n(\theta_1)-Q_0(\theta_1)|>\frac{\kappa}{2}\Big) \\ &\leq \mathbb{P}\Big(\{Q_0(\hat{\theta}_1)\geq\kappa\}\cap\big\{\sup_{\theta_1\in\Theta_1}|Q_n(\theta_1)-Q_0(\theta_1)|\leq\frac{\kappa}{2}\big\}\Big)+\mathbb{P}\Big(\sup_{\theta_1\in\Theta_1}|Q_n(\theta_1)-Q_0(\theta_1)|>\frac{\kappa}{2}\Big). \end{aligned}$$

For the first probability on the RHS, by set inclusion, $\{Q_0(\hat{\theta}_1) \geq \kappa\} \cap \{\sup_{\theta_1 \in \Theta_1} |Q_n(\theta_1) - Q_0(\theta_1)| \leq \kappa/2\} \subseteq \{Q_n(\hat{\theta}_1) \geq \kappa/2\}$. Therefore,

$$\begin{aligned}\mathbb{P}\Big(\{Q_0(\hat{\theta}_1) \geq \kappa\} \cap \Big\{ \sup_{\theta_1 \in \Theta_1} |Q_n(\theta_1) - Q_0(\theta_1)| \leq \frac{\kappa}{2} \Big\}\Big) &\leq \mathbb{P}\Big(Q_n(\hat{\theta}_1) \geq \frac{\kappa}{2}\Big) \\ &\leq o(1),\end{aligned}$$

and

$$\mathbb{P}(Q_0(\hat{\theta}_1) \geq \kappa) \leq \mathbb{P}\Big(\sup_{\theta_1 \in \Theta_1} |Q_n(\theta_1) - Q_0(\theta_1)| > \frac{\kappa}{2}\Big) + o(1).$$

Note that data are i.i.d., Θ_1 is compact, for all $z \in \mathcal{Z}$, $g_1(z, \theta_1, h_{10})$ is continuous at each $\theta_1 \in \Theta_1$ w.p.1, and $||g_1(z, \theta_1, h_{10})|| \leq d_1(z)$ for all $\theta_1 \in \Theta_1$ where $\mathbb{E}[d_1(z)] < \infty$ by Assumption 4.3(i), the hypotheses of Lemma 2.4 of Newey and McFadden [26], p.2129, are satisfied. Hence, by UWL,

$$\mathbb{P}\Big(\sup_{\theta_1 \in \Theta_1} |Q_n(\theta_1) - Q_0(\theta_1)| > \frac{\kappa}{2}\Big) = o(1).$$

Hence, $\mathbb{P}(Q_0(\hat{\theta}_1) \geq \kappa) = o(1)$. By $(A.4)$, $\mathbb{P}(||\hat{\theta}_1 - \theta_{10}|| > \epsilon) \leq o(1)$ for all $\epsilon > 0$. That is, $\hat{\theta}_1$ is consistent. □

Proof of Theorem 4.1(ii): Asymptotic Normality

The estimator of $\hat{\theta}_1$ of θ_{10} solves

$$\sum_{i=1}^{n} \hat{\pi}_i \hat{g}_{1i} = 0.$$

By a Taylor expansion around $\hat{\theta}_1 = \theta_{10}$, for some $\bar{\theta}_1$ on the line segment joining $\hat{\theta}_1$ and θ_{10},

$$\sum_{i=1}^{n} \hat{\pi}_i g_{1i}(\theta_{10}, \hat{h}_1) + \sum_{i=1}^{n} \hat{\pi}_i G_{1i}(\bar{\theta}_1, \hat{h}_1)(\hat{\theta}_1 - \theta_{10}) = 0. \tag{A.5}$$

By Lemma B.6,

$$\begin{aligned}\sum_{i=1}^{n} \hat{\pi}_i G_{1i}(\bar{\theta}_1, \hat{h}_1) &= \frac{1}{n}\sum_{i=1}^{n} G_{1i}(\bar{\theta}_1, \hat{h}_1) + \Big(\frac{1}{n}\sum_{i=1}^{n} G_{1i}(\bar{\theta}_1, \hat{h}_1)\hat{g}'_{2i}\Big)\hat{\lambda}(1 + o_p(1)) \\ &\quad + \Big(\frac{1}{n}\sum_{i=1}^{n} G_{1i}(\bar{\theta}_1, \hat{h}_1)\Big) O_p(n^{-1}) \\ &:= L1 + L2 + L3.\end{aligned}$$

Note that

$$\begin{aligned}\Big|\Big|\frac{1}{n}\sum_{i=1}^{n}[G_{1i}(\bar{\theta}_1,\hat{h}_1)-G_{1i}]\Big|\Big| &\leq \Big(\frac{1}{n}\sum_{i=1}^{n}\tilde{b}_1(z_i)\Big)\big(||\bar{\theta}_1-\theta_{10}||+||\hat{h}_1-h_{10}||\big)\\ &= \big(\mathbb{E}[\tilde{b}_1(z)]+o_p(1)\big)\big(||\bar{\theta}_1-\theta_{10}||+||\hat{h}_1-h_{10}||\big)\\ &= o_p(n^{-\frac{1}{4}}),\end{aligned}$$

where the inequality follows by CS and Assumption 4.3(iv), the first equality by WLLN, and the second equality follows by Assumption 4.3(iv), $||\hat{h}_1-h_{10}||=o_p(n^{-\frac{1}{4}})$ and $||\hat{\theta}_1-\theta_{10}||=O_p(n^{-\frac{1}{2}})$ by Lemma B.5.

Hence, by T, CS, and WLLN, $L1=G_1+o_p(1)$.

For L2,

$$\begin{aligned}\Big|\Big|\frac{1}{n}\sum_{i=1}^{n}G_{1i}(\bar{\theta}_1,\hat{h}_1)\hat{g}_{2i}'\Big|\Big| &\leq \frac{1}{n}\sum_{i=1}^{n}\tilde{d}_1(z_i)d_2(z_i)\\ &= O_p(1),\end{aligned}$$

by CS, Assumptions 4.3(i), (ii), 4.4(i) and M.

Hence, by CS, and $||\hat{\lambda}||=O_p(n^{-\frac{1}{2}})$ by Lemma B.2, $||L_2||=O_p(n^{-\frac{1}{2}})$.

Also, by CS, Assumption 4.3(ii), and M, $||L_3||=O_p(n^{-1})$.

Therefore, by the above arguments,

$$\sum_{i=1}^{n}\hat{\pi}_i G_{1i}(\bar{\theta}_1,\hat{h}_1)=G_1+o_p(1). \tag{A.6}$$

Using Lemma B.6,

$$\begin{aligned}\sum_{i=1}^{n}\hat{\pi}_i g_{1i}(\theta_{10},\hat{h}_1) &= \frac{1}{n}\sum_{i=1}^{n}g_{1i}(\theta_{10},\hat{h}_1)+\Big(\frac{1}{n}\sum_{i=1}^{n}g_{1i}(\theta_{10},\hat{h}_1)\hat{g}_{2i}'\Big)\hat{\lambda}(1+o_p(1))\\ &\quad+\Big(\frac{1}{n}\sum_{i=1}^{n}g_{1i}(\theta_{10},\hat{h}_1)\Big)O_p(n^{-1}).\end{aligned}$$

Note that,

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{g}_{1i}^h\hat{g}_{2i}'-g_{1i}g_{2i}')=\frac{1}{n}\sum_{i=1}^{n}(\hat{g}_{1i}^h\hat{g}_{2i}'-\hat{g}_{1i}^hg_{2i}')+\frac{1}{n}\sum_{i=1}^{n}(\hat{g}_{1i}^hg_{2i}'-g_{1i}g_{2i}').$$

Now,

$$\begin{aligned}\Big|\Big|\frac{1}{n}\sum_{i=1}^{n}\hat{g}_{1i}^{h}(\hat{g}_{2i}-g_{2i})'\Big|\Big| &\leq \Big(\frac{1}{n}\sum_{i=1}^{n}d_1(z_i)b_2(z_i)\Big)\big(||\hat{\theta}_2-\theta_{20}||+||\hat{h}_2-h_{20}||\big)\\ &= \big(\mathbb{E}[d_1(z)b_2(z)]+o_p(1)\big)\big(||\hat{\theta}_2-\theta_{20}||+||\hat{h}_2-h_{20}||\big)\\ &= o_p(n^{-\frac{1}{4}}),\end{aligned}$$

where the inequality follows by CS and Assumptions 4.3(i), (iii), the first equality by WLLN, and the second equality follows by Assumptions 4.3(i), (iii), 4.4(i), $||\hat{\theta}_2-\theta_{20}||=O_p(n^{-\frac{1}{2}})$ by Lemma B.5, and $||\hat{h}_2-h_{20}||=o_p(n^{-\frac{1}{4}})$.

Similarly,

$$\begin{aligned}\Big|\Big|\frac{1}{n}\sum_{i=1}^{n}(\hat{g}_{1i}^{h}-g_{1i})g_{2i}'\Big|\Big| &\leq \Big(\frac{1}{n}\sum_{i=1}^{n}b_1(z_i)d_2(z_i)\Big)||\hat{h}_1-h_{10}||\\ &= \big(\mathbb{E}[b_1(z)d_2(z)]+o_p(1)\big)||\hat{h}_1-h_{10}||\\ &= o_p(n^{-\frac{1}{4}}),\end{aligned}$$

where the inequality follows by CS and Assumptions 4.3(i), (iii), the first equality by WLLN, and the second equality follows by Assumptions 4.3(i), (iii), 4.4(i) and $||\hat{h}_1-h_{10}||=o_p(n^{-\frac{1}{4}})$.

Hence, by CS, T, $\hat{\lambda}=O_p(n^{-\frac{1}{2}})$ by Lemma B.2, and WLLN,

$$\Big(\frac{1}{n}\sum_{i=1}^{n}\hat{g}_{1i}^{h}\hat{g}_{2i}'\Big)\hat{\lambda}(1+o_p(1))=B\hat{\lambda}+o_p(n^{-\frac{1}{2}}).$$

Also, $\sum_{i=1}^{n}\hat{g}_{1i}^{h}=O_p(n^{-\frac{1}{2}})$ by Lemma 5.1 of Newey [25]. Thus,

$$\sum_{i=1}^{n}\hat{\pi}_i\hat{g}_{1i}^{h}=\frac{1}{n}\sum_{i=1}^{n}\hat{g}_{1i}^{h}+B\hat{\lambda}+o_p(n^{-\frac{1}{2}}). \tag{A.7}$$

By Assumption 4.1(ii), G_1 has full column rank. Using this, $(A.5)$, $(A.6)$ and $(A.7)$,

$$\sqrt{n}(\hat{\theta}_1-\theta_{10})=-G_1^{-1}\Big[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{g}_{1i}^{h}+B\sqrt{n}\hat{\lambda}+o_p(1)\Big].$$

Therefore, by Lemma B.5,

$$\sqrt{n}(\hat{\theta}_1-\theta_{10})=-G_1^{-1}\Big[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{g}_{1i}^{h}-BP\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{g}_{2i}^{h}\Big]+o_p(1).$$

By CLT,

$$G_1\sqrt{n}(\hat{\theta}_1-\theta_{10})\xrightarrow{d}\mathcal{N}(0,\Sigma),$$

where

$$\begin{aligned}\Sigma &= \begin{bmatrix} I & -BP \end{bmatrix} \begin{bmatrix} V_1 & B^\star \\ B^{\star\prime} & V_2 \end{bmatrix} \begin{bmatrix} I \\ -PB' \end{bmatrix} \\ &= V_1 - BPB^\star - B^\star PB + BPV_2PB'.\end{aligned}$$

The result follows by Cramer's theorem. □

Proof of Theorem 4.1(iii): Variance Structure for Locally Robust Estimation

By Assumption 4.4(ii), the same arguments as for the proof for part (i) with $\psi_j(z, \theta_j, h_j)$ replacing $g_j(z, \theta_j, h_j)$ $(j = 1, 2)$ hold, so that

$$\sqrt{n}(\hat{\theta}_1 - \theta_{10}) = -G_1^{\star -1}\Big[\frac{1}{\sqrt{n}}\sum_{i=1}^n \hat{\psi}_{1i}^h - B^\star P^\star \frac{1}{\sqrt{n}}\sum_{i=1}^n \hat{\psi}_{2i}^h\Big] + o_p(1).$$

Thus, by CLT,

$$G_1^{\star -1}\sqrt{n}(\hat{\theta}_1 - \theta_{10}) \xrightarrow{d} \mathcal{N}(0, \Sigma^\star),$$

where

$$\begin{aligned}\Sigma^\star &= [I, -B^\star P^\star] \begin{bmatrix} V_1 & B^\star \\ B^{\star\prime} & V_2 \end{bmatrix} \begin{bmatrix} I \\ -P^\star B^{\star\prime} \end{bmatrix} \\ &= V_1 - B^\star P^\star B^{\star\prime},\end{aligned}$$

since $P^\star V_2 P^\star = P^\star$. The result follows by Cramer's theorem. □

A.1 Proof of Corollary 5.1(i): Double Robustness

Under general misspecification, assume the propensity score estimates and conditional expectation function estimates converge to psuedo-true values; $\hat{p}(x) \xrightarrow{p} p_\star(x)$, $\hat{q}(x) := \hat{q}(x; \tilde{\theta}_1) \xrightarrow{p} q_\star(x)$, where $\tilde{\theta}_1$ is a preliminary estimator for θ_{10} obtained, for example, by the Horvitz and Thompson [23] inverse probability weighting method. The auxiliary moment restriction $\mathbb{E}[g_2(d, z, \theta_{20}, h_{20}(x))] = 0$ is assumed to be correctly specified. Let $g_{1i}(\theta_1) = g_1(z_i, \theta_1)$ $(i = 1, ..., n)$. From equation (5.7), the estimator $\hat{\theta}_1$ solves

$$\sum_{i=1}^n \hat{\pi}_i\Big(\frac{d_i}{\hat{p}(x_i)} g_{1i}(\hat{\theta}_1) - \Big(\frac{d_i}{\hat{p}(x_i)} - 1\Big)\hat{q}(x_i)\Big) = 0.$$

By Assumptions 5.1, 5.2, and Lemma B.6,

$$\begin{aligned} 0 &= \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{d_i}{\hat{p}(x_i)}g_{1i}(\hat{\theta}_1) - \Big(\frac{d_i}{\hat{p}(x_i)} - 1\Big)\hat{q}(x_i)\Big) + \Big(\frac{1}{n}\sum_{i=1}^{n}\frac{d_i}{\hat{p}(x_i)}g_{1i}(\hat{\theta}_1)g_{2i}(\hat{\theta}_2,\hat{h}_2)'\Big)\hat{\lambda} \\ &\quad -\Big[\frac{1}{n}\sum\Big(\frac{d_i}{\hat{p}(x_i)} - 1\Big)\hat{q}(x_i)g_{2,i}(\hat{\theta}_2,\hat{h}_2)'\Big]\hat{\lambda} + o_p(1). \end{aligned}$$

By the proof of Theorem 4.1,

$$\frac{1}{n}\sum_{i=1}^{n}\Big(\frac{d_i}{\hat{p}(x_i)}g_{1i}(\hat{\theta}) - \Big(\frac{d_i}{\hat{p}(x_i)} - 1\Big)\hat{q}(x_i)\Big)g_{2i}(\hat{\theta}_2,\hat{h}_2)' = O_p(1).$$

Hence, since $\hat{\lambda} = O_p(n^{-\frac{1}{2}})$ by Lemma B.2, by CS,

$$\sum_{i=1}^{n}\hat{\pi}_i\Big(\frac{d_i}{\hat{p}(x_i)}g_{1i}(\hat{\theta}_1) - \Big(\frac{d_i}{\hat{p}(x_i)} - 1\Big)\hat{q}(x_i)\Big) = \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{d_i}{\hat{p}(x_i)}g_{1i}(\hat{\theta}_1) - \Big(\frac{d_i}{\hat{p}(x_i)} - 1\Big)\hat{q}(x_i)\Big) + o_p(1).$$

It is now shown that if $\hat{\theta}_1$ satisfies

$$\frac{1}{n}\sum_{i=1}^{n}\Big(\frac{d_i}{\hat{p}(x_i)}g_{1i}(\hat{\theta}_1) - \Big(\frac{d_i}{\hat{p}(x_i)} - 1\Big)\hat{q}(x_i)\Big) = o_p(1), \tag{A.8}$$

then $\hat{\theta}_1$ is consistent for θ_{10} if at least one of $p_\star(x) = p_0(x)$ and $q_\star(x) = q_0(x,\theta_{10})$ holds.

Condition A.1 ($p_\star(x) \neq \mathbb{P}(d=1|x)$ and $q_\star(x) = q_0(x;\theta_{10})$)**.** By WLLN, continuity of the expectation at $p_\star(x)$ and $q_0(x;\theta_{10})$, the $\hat{\theta}_1$ that satisfies $(A.8)$ is consistent for the θ_1 that satisfies

$$\mathbb{E}\Big[\frac{d}{p_\star(x)}g_1(z,\theta_1) - \Big(\frac{d}{p_\star(x)} - 1\Big)q_0(x;\theta_{10})\Big] = o(1),$$

or,

$$\mathbb{E}\Big[\frac{d}{p_\star(x)}\big(g_1(z,\theta_1) - q_0(x;\theta_{10})\big) + q_0(x;\theta_{10})\Big] = o(1).$$

Hence,

$$\mathbb{E}\Big[\frac{d}{p_\star(x)}\big(g_1(z,\theta_1) - q_0(x;\theta_{10})\big)\Big] = o(1)$$

since $\mathbb{E}[q_0(x;\theta_{10})] = \mathbb{E}[\mathbb{E}[g_1(z,\theta_{10})|x]] = \mathbb{E}[g_1(z,\theta_{10})] = 0$ by LIE. By MAR,

$$\mathbb{E}\Big[\frac{\mathbb{E}[d|x]}{p_\star(x)}\mathbb{E}[g_1(z,\theta_1) - g_1(z,\theta_{10})|x]\Big] = o(1).$$

Therefore, by continuity of g_1, and uniqueness of the true value θ_{10} for $\mathbb{E}[g_1(z,\theta_1)] = 0$, $\theta_1 = \theta_{10}$ w.p.a. 1. See Wooldridge ([40], pp.1296-7) and Graham et al. ([18], p.1073) for similar arguments.

Condition A.2 ($p_\star(x) = p_0(x)$ and $q_\star(x) \neq q_0(x;\theta_{10})$)**.** By WLLN, continuity of the expec-

tation at $p_0(x)$ and $q_\star(x)$, the $\hat{\theta}_1$ that satisfies $(A.8)$ is consistent for the θ_1 that satisfies

$$\mathbb{E}\Big[\frac{d}{p_0(x)}g_1(z,\theta_1) - \Big(\frac{d}{p_0(x)} - 1\Big)q_\star(x)\Big] = o(1)$$

Now,

$$\mathbb{E}\Big[\frac{d}{p_0(x)}g_1(z,\theta_1)\Big] - \mathbb{E}\Big[\Big(\frac{\mathbb{E}[d|x]}{p_0(x)} - 1\Big)q_\star(x)\Big] = o(1).$$

Hence, by LIE and $\mathbb{E}[d|x] = p_0(x)$,

$$\mathbb{E}\Big[\frac{d}{p_0(x)}g_1(z,\theta_1)\Big] = o(1).$$

Then,

$$\mathbb{E}\Big[\frac{\mathbb{E}[d|y,x]}{p_0(x)}g_1(z,\theta_1)\Big] = o(1)$$

by LIE. Finally, by MAR and $\mathbb{E}[d|x] = p_0(x)$,

$$\mathbb{E}[g_1(z,\theta_1)] = o(1).$$

Finally, $\mathbb{E}[g_1(z,\theta_1)] = o_p(1)$ implies that $\theta_1 = \theta_{10}$ by Assumption 5.1(i). □

Proof of Corollary 5.1(ii): Asymptotic Variance

Apply Theorem 4.1(iii), with ψ_j there equal $\tilde{g}_j$ $(j = 1, 2)$, and h_1 there equal to $(p, q)'$. This leads to the asymptotic variance of $\hat{\theta}_1$ being equal to $G_1^{-1}(\Sigma_0 - \Sigma_1 - \Sigma_2)G_1'^{-1}$. □

B GEL Lemmata

Lemma B.1 (Newey and Smith [28], Lemma A1, with nuisance estimation)**.** *Suppose $\bar{h}_2$ is an estimator for h_{20} such that $||\bar{h}_2 - h_{20}|| = o_p(1)$. Under Assumptions 4.1-4.3, for $1/\alpha < \zeta < 1/2$,*

$$\max_{1\le i\le n}\sup_{\theta_2\in\Theta_2}\big|\lambda' g_2(z_i,\theta_2,\bar{h}_2)\big| \xrightarrow{p} 0.$$

Proof. Since $||\bar{h}_2 - h_{20}|| = o_p(1)$, $||\bar{h}_2 - h_{20}|| < \epsilon$ for any $\epsilon > 0$ w.p.1. Then,

$$\begin{aligned}
\max_{1\le i\le n}\sup_{\theta_2\in\Theta_2,\lambda\in\Lambda_n}\big|\lambda' g_2(z_i,\theta_2,\bar{h}_2)\big| &\le \sup_{\lambda\in\Lambda_n}||\lambda|| \times \max_{1\le i\le n}\sup_{\theta_2\in\Theta_2}||g_2(z_i,\theta_2,\bar{h}_2)|| \\
&\le n^{-\zeta} d_2(z_i) \\
&= O_p(n^{-\zeta+\frac{1}{\alpha}}) \\
&= o_p(1),
\end{aligned}$$

where the first inequality follows from CS. The second inquality follows from Assumption 4.3(i) and definition of the set Λ_n. The first equality follows by noting that by Assumption 4.3(i), $\mathbb{E}[d_2(z)^\alpha] < \infty$, so by M, $\max_{1\leq i\leq n} d_2(z_i) \leq O_p(n^{\frac{1}{\alpha}})$. The second equality follows from $1/\alpha < \zeta$. □

Lemma B.2 (Newey and Smith [28], Lemma A2, with nuisance estimation)**.** *Under Assumptions 4.1-4.4, if* $\bar{\theta}_2 \in \Theta_2$, $\bar{\theta}_2 \xrightarrow{p} \theta_{20}$, *and* $\sum_{i=1}^n g_2(z_i, \bar{\theta}_2, \hat{h}_2)/n = O_p(n^{-\frac{1}{2}})$, *then* $\bar{\lambda} = \arg\max_{\lambda\in\hat{\Lambda}_n(\bar{\theta}_2,\hat{h}_2)} \hat{P}_n(\bar{\theta}_2, \hat{h}_2, \lambda)$ *exists w.p.a.1,* $\bar{\lambda} = O_p(n^{-\frac{1}{2}})$, *and* $\sup_{\lambda\in\hat{\Lambda}_n(\bar{\theta}_2,\hat{h}_2)} \hat{P}_n(\bar{\theta}_2, \hat{h}_2, \lambda) \leq \rho_0 + O_p(n^{-1})$.

Proof. Let $\bar{g}_{2i} = g_2(z_i, \bar{\theta}_2, \hat{h}_2)$. By T,

$$\begin{aligned}\Big|\Big|\frac{1}{n}\sum_{i=1}^n(\bar{g}_{2i}\bar{g}_{2i}' - g_{2i}g_{2i}')\Big|\Big| &\leq \Big|\Big|\frac{1}{n}\sum_{i=1}^n g_{2i}(\bar{g}_{2i} - g_{2i})'\Big|\Big| + \Big|\Big|\frac{1}{n}\sum_{i=1}^n(\bar{g}_{2i} - g_{2i})g_{2i}'\Big|\Big| \\ &\quad + \Big|\Big|\frac{1}{n}\sum_{i=1}^n(\bar{g}_{2i} - g_{2i})(\bar{g}_{2i} - g_{2i})'\Big|\Big| \\ &= E1 + E2 + E3.\end{aligned}$$

Now,

$$\begin{aligned}E1 &\leq \Big(\frac{1}{n}\sum_{i=1}^n d_2(z_i)b_2(z_i)\Big)\big(||\bar{\theta}_2 - \theta_{20}|| + ||\hat{h}_2 - h_{20}||\big) \\ &= \big(\mathbb{E}[d_2(z)b_2(z)] + o_p(1)\big)\big(||\bar{\theta}_2 - \theta_{20}|| + ||\hat{h}_2 - h_{20}||\big) \\ &\leq o_p(1),\end{aligned}$$

by CS, M, Assumptions 4.3(i), (iii), 4.4(i), and consistency of $\bar{\theta}_2$ and $\hat{h}_2$.

$E2 \leq o_p(1)$ by identical arguments.

$$\begin{aligned}E3 &\leq \Big(\frac{1}{n}\sum_{i=1}^n b_2^2(z_i)\Big)\big(||\bar{\theta}_2 - \theta_{20}|| + ||\hat{h} - h_0||\big)^2 \\ &= \big(\mathbb{E}[b_2^2(z)] + o_p(1)\big)\big(||\bar{\theta}_2 - \theta_{20}|| + ||\hat{h} - h_0||\big)^2 \\ &\leq o_p(1),\end{aligned}$$

by CS, M, Assumption 4.4(iii), and consistency of $\bar{\theta}_2$ and $\hat{h}_2$.

By the above results, T and WLLN,

$$\Big|\Big|\frac{1}{n}\sum_{i=1}^n \bar{g}_{2i}\bar{g}_{2i}' - \Omega_2\Big|\Big| = o_p(1).$$

(The rest of the proof is identical to the proof of Lemma A2 of Newey and Smith [28]).

Thus, by nonsingularity of Ω_2 the smallest eigenvalue of $\bar{\Omega}_2 = \sum_{j=1}^n \bar{g}_{2j}\bar{g}_{2j}'/n$ is bounded away from zero w.p.a.1. Let Λ_n be as defined in Section 2.2. By Lemma B.1 and twice continuous differentiability of $\rho(v)$ in a neighborhood of zero, $\hat{P}_n(\bar{\theta}_2, \hat{h}_2, \lambda)$ is twice continuously differentiable on Λ_n w.p.a.1. Then $\tilde{\lambda} = \arg\max_{\lambda\in\Lambda_n} \hat{P}_n(\bar{\theta}_2, \hat{h}_2, \lambda)$ exists w.p.a.1. Furthermore, for $\bar{g}_{2i} = g_2(z_i, \bar{\theta}_2, \hat{h}_2)$ and any $\dot{\lambda}$ on a line segment joining $\tilde{\lambda}$ and 0, by Lemma B.1 and $\rho_2 = -1$, $\max_{1\le i\le n} \rho_2(\dot{\lambda}'\bar{g}_{2i}) < -1/2$ w.p.a.1. Then by a Taylor expansion around $\lambda = 0$, there is a $\dot{\lambda}$ on the line segment joining $\tilde{\lambda}$ and 0 such that for $\bar{g}_2 = \sum_{i=1}^n g_2(z_i, \bar{\theta}_2, \hat{h}_2)/n$,

$$\begin{aligned}\rho_0 &= P_n(\bar{\theta}_2, \hat{h}_2, 0) \le P_n(\bar{\theta}_2, \hat{h}_2, \tilde{\lambda}) = \rho_0 - \tilde{\lambda}'\bar{g}_2 + \frac{1}{2}\tilde{\lambda}'\Big(\frac{1}{n}\sum_{i=1}^n \rho_2(\dot{\lambda}'\bar{g}_{2i})\bar{g}_{2i}\bar{g}_{2i}'\Big)\tilde{\lambda} \\ &\le \rho_0 - \tilde{\lambda}'\bar{g}_2 - \frac{1}{4}\tilde{\lambda}'\bar{\Omega}_2\tilde{\lambda} \le \rho_0 + ||\tilde{\lambda}||\,||\bar{g}_2|| - C||\tilde{\lambda}||^2.\end{aligned}$$

Substracting $\rho_0 - C||\tilde{\lambda}||^2$ from both sides and dividing by $||\tilde{\lambda}||$, $C||\tilde{\lambda}|| \le ||\bar{g}_2||$, w.p.a.1. Since $\bar{g}_2 = O_p(n^{-\frac{1}{2}})$ by assumption, $||\tilde{\lambda}|| = O_p(n^{-\frac{1}{2}}) = o_p(n^{-\zeta})$. Therefore, w.p.a.1 $\tilde{\lambda} \in \text{int}(\Lambda_n)$ and hence $\partial\hat{P}_n(\bar{\theta}_2, \tilde{\lambda})/\partial\lambda = 0$, the first order conditions for an interior maximum. By Lemma B.1, w.p.a.1, $\tilde{\lambda} \in \hat{\Lambda}_n(\bar{\theta}_2, \hat{h}_2)$, so by concavity of $\hat{P}_n(\bar{\theta}_2, \hat{h}_2, \lambda)$ and convexity of $\hat{\Lambda}_n(\bar{\theta}_2, \hat{h}_2)$, it follows that $\hat{P}_n(\bar{\theta}_2, \hat{h}_2, \tilde{\lambda}) = \sup_{\lambda\in\hat{\Lambda}_n(\bar{\theta}_2,\hat{h}_2)} \hat{P}_n(\bar{\theta}_2, \hat{h}_2, \lambda)$, giving the first and second conclusions with $\bar{\lambda} = \tilde{\lambda}$. Then by the last inequality of the above equation, $||\bar{g}_2|| = O_p(n^{-\frac{1}{2}})$, and $||\tilde{\lambda}|| = O_p(n^{-\frac{1}{2}})$, $\hat{P}_n(\bar{\theta}_2, \hat{h}_2, \bar{\lambda}) \le \rho_0 + ||\bar{\lambda}||\,||\bar{g}_2|| - C||\bar{\lambda}||^2 = \rho_0 + O_p(n^{-1})$. □

Lemma B.3 (Newey and Smith [28], Lemma A3, with nuisance estimation)**.** *Under Assumptions 4.1,4.2, 4.3(i),* $||\hat{g}_2|| = O_p(n^{-\frac{1}{2}})$.

Proof. Recall $\hat{g}_{2i} = g_{2i}(\hat{\theta}_2, \hat{h}_2)$, $\hat{g}_2 = \sum_{i=1}^n g_{2i}(\hat{\theta}_2, \hat{h}_2)/n$, and for ζ in Section 2.2, $\tilde{\lambda} = -n^{-\zeta}\hat{g}_2/||\hat{g}_2||$. By Lemma B.1, $\max_{i\le n} |\tilde{\lambda}'\hat{g}_{2i}| \xrightarrow{p} 0$ and $\tilde{\lambda} \in \hat{\Lambda}_n(\hat{\theta}_2)$ w.p.a.1. Thus, for any $\dot{\lambda}$ on the line segment joining $\tilde{\lambda}$ and 0, w.p.a.1 $\rho_2(\dot{\lambda}'\hat{g}_{2i}) \ge -C$ $(i = 1, ..., n)$ for some $0 < C < 1$. Also, by CS and Assumption 4.3(i), $\sum_{i=1}^n \hat{g}_{2i}\hat{g}_{2i}'/n \le (\sum_{i=1}^n d_2(z_i)^2/n)I \xrightarrow{p} CI$, for some $C > 0$, so that the largest eigenvalue of $\sum_{i=1}^n \hat{g}_{2i}\hat{g}_{2i}'/n$ is bounded above w.p.a.1. A Taylor expansion around 0 then gives

$$\begin{aligned}\hat{P}_n(\hat{\theta}_2, \hat{h}_2, \tilde{\lambda}) &= \rho_0 - \tilde{\lambda}'\hat{g}_2 + \frac{1}{2}\tilde{\lambda}'\Big(\frac{1}{n}\sum_{i=1}^n \rho_2(\dot{\lambda}'\hat{g}_{2i})\hat{g}_{2i}\hat{g}_{2i}'\Big)\tilde{\lambda} \\ &\ge \rho_0 + n^{-\zeta}||\hat{g}_2|| - \frac{C}{2}\tilde{\lambda}'\Big(\frac{1}{n}\sum_{i=1}^n \hat{g}_{2i}\hat{g}_{2i}'\Big)\tilde{\lambda} \ge \rho_0 + n^{-\zeta}||\hat{g}_2|| - Cn^{-2\zeta}\end{aligned}$$

for some $C > 0$ w.p.a.1, using $\tilde{\lambda} \in \hat{\Lambda}_n(\hat{\theta}_2, \hat{h}_2)$ w.p.a.1. By Lemma 5.1 of Newey [25], p.1366, $\hat{g}(\theta_{20}, \hat{h}_2) = O_p(n^{-\frac{1}{2}})$. Thus the hypotheses of Lemma B.2. are satisfied by $\bar{\theta}_2 = \theta_{20}$. Now,

$$\rho_0 + n^{-\zeta}||\hat{g}_2|| - Cn^{-2\zeta} \le \hat{P}_n(\hat{\theta}_2, \hat{h}_2, \tilde{\lambda}) \le \hat{P}_n(\hat{\theta}_2, \hat{h}_2, \hat{\lambda}) \le \sup_{\lambda\in\hat{\Lambda}_n(\theta_{20},\hat{h}_2)} \hat{P}_n(\theta_{20}, \hat{h}_2, \lambda) \le \rho_0 + O_p(n^{-1}),$$

where the first inequality follows by the above equation, the second inequality follows by the definition of $\hat{\lambda}$ being a maximiser, the third inequality follows from $\hat{\theta}_2$ being a minimiser, and the fourth inequality follows from Lemma B.2. (The rest of the proof is identical to the proof of Lemma A3 of Newey and Smith [28]).

Also, by $\zeta < 1/2$, it follows that $\zeta - 1 < -1/2 < -\zeta$. Solving the above equation for $||\hat{g}_2||$ gives

$$||\hat{g}_2|| \leq Cn^{-\zeta} + O_p(n^{\zeta-1}) = O_p(n^{-\zeta}).$$

Now consider any $\varepsilon_n \to 0$. Let $\bar{\lambda} = -\varepsilon_n \hat{g}_2$. Note that $\bar{\lambda} = o_p(n^{-\zeta})$ by the above, so that $\bar{\lambda} \in \Lambda_n$, w.p.a.1. Then, as in the second last equation,

$$\rho_0 + \bar{\lambda}'\hat{g}_2 - C||\bar{\lambda}||^2 = \rho_0 + \varepsilon_n||\hat{g}_2||^2 - C\varepsilon_n^2||\hat{g}_2||^2 \leq \rho_0 + O_p(n^{-1})$$

where the first equality follows by definition of $\bar{\lambda}$, and the second follows from above arguments. Since, for all n large enough, $1 - \varepsilon_n C$ is bounded away from zero, it follows that $\varepsilon_n||\hat{g}_2||^2 = O_p(n^{-1})$. The conclusion then follows by a result, that if $\varepsilon_n Y_n = O_p(n^{-1})$ for all $\varepsilon_n \to 0$, then $Y_n = O_p(n^{-1})$. □

Lemma B.4. *Under Assumptions 4.1, 4.2, 4.3(i), (iii),* $\hat{\theta}_2 \xrightarrow{p} \theta_{20}$.

Proof. By Lemma B.3, $\sum_{i=1}^n g_{2i}(\hat{\theta}_2, \hat{h}_2)/n \xrightarrow{p} 0$. Now,

$$\begin{aligned}
\Big|\Big|\frac{1}{n}\sum_{i=1}^n g_{2i}(\hat{\theta}_2, \hat{h}_2) - \mathbb{E}[g_2(z, \hat{\theta}_2, h_{20})]\Big|\Big| &\leq \Big|\Big|\frac{1}{n}\sum_{i=1}^n \big(g_{2i}(\hat{\theta}_2, \hat{h}_2) - g_{2i}(\hat{\theta}_2, h_{20})\big)\Big|\Big| \\
&\quad + \Big|\Big|\frac{1}{n}\sum_{i=1}^n g_{2i}(\hat{\theta}_2, h_{20}) - \mathbb{E}[g_2(z, \hat{\theta}_2, h_{20})]\Big|\Big| \\
&:= H1 + H2
\end{aligned}$$

Note that

$$\begin{aligned}
H1 &\leq \Big(\frac{1}{n}\sum_{i=1}^n b_2(z_i)\Big)||\hat{h}_2 - h_{20}|| \\
&= \big(\mathbb{E}[b_2(z)] + o_p(1)\big)||\hat{h}_2 - h_{20}|| \\
&\leq o_p(1)
\end{aligned}$$

where the first inequality follows by CS and Assumption 4.3(iii), the equality by WLLN, and the second inequality follows by Assumption 4.3(iii) and consistency of $\hat{h}_2$.

$H2 \leq o_p(1)$ by UWL. By T, $\mathbb{E}[g_2(z, \hat{\theta}_2, h_{20})] \xrightarrow{p} 0$. Since $\mathbb{E}[g_2(z, \theta_2, h_{20})]$ is uniquely zero at $\theta_2 = \theta_{20}$ and $g_2(\theta_2, h_{20})$ is continuous in $\theta_2 \in \Theta_2$, $||\mathbb{E}[g_2(z, \theta_2, h_{20})]||$ must be bounded away from zero outside any neighborhood of θ_{20}. Therefore, $\hat{\theta}_2$ must be inside any neighborhood of θ_{20} w.p.a.1, that is, $\hat{\theta}_2 \xrightarrow{p} \theta_{20}$. □

Lemma B.5 (***GEL Estimation***). *Under Assumptions 4.1-4.4, the GEL estimator $\hat{\theta}_2$ and the GEL Lagrange multplier $\hat{\lambda}$ satisfies*

$$\begin{aligned}\sqrt{n}(\hat{\theta}_2 - \theta_{20}) &= -(G_2'\Omega_2 G_2)^{-1} G_2'\Omega_2^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n g_{2i}(\theta_{20}, \hat{h}_2) + o_p(1) \\ \sqrt{n}(\hat{\lambda} - 0) &= -P \frac{1}{\sqrt{n}} \sum_{i=1}^n g_{2i}(\theta_{20}, \hat{h}_2) + o_p(1).\end{aligned}$$

Proof. The first order conditions from GEL estimation are

$$\frac{1}{n} \sum_{i=1}^n \rho_1(\hat{\lambda}'\hat{g}_{2i}) \hat{g}_{2i} = 0 \tag{B.1}$$

$$\frac{1}{n} \sum_{i=1}^n \rho_1(\hat{\lambda}'\hat{g}_{2i}) \hat{G}_{2i}' \hat{\lambda} = 0 \tag{B.2}$$

By Taylor expansion around 0, for some $\bar{\lambda}$ on the line segment joining $\hat{\lambda}$ and 0,

$$\rho_1(\hat{\lambda}'\hat{g}_{2i}) = -1 + \rho_2(\bar{\lambda}'\hat{g}_{2i}) \hat{\lambda}'\hat{g}_{2i}, \ (i = 1, ..., n).$$

Therefore,

$$\frac{1}{n} \sum_{i=1}^n \rho_2(\hat{\lambda}'\hat{g}_{2i}) \hat{g}_{2i} = -\frac{1}{n} \sum_{i=1}^n \hat{g}_{2i} + \Big(\frac{1}{n} \sum_{i=1}^n \rho_2(\bar{\lambda}'\hat{g}_{2i}) \hat{g}_{2i} \hat{g}_{2i}' \Big) \hat{\lambda}.$$

By Lemma B.1, $\rho_2(\bar{\lambda}'\hat{g}_{2i}) = -1 + o_p(1)$, uniformly $(i = 1, ..., n)$. Hence,

$$\frac{1}{n} \sum_{i=1}^n \rho_2(\bar{\lambda}'\hat{g}_{2i}) \hat{g}_{2i} \hat{g}_{2i}' = -\frac{1}{n} \sum_{i=1}^n \hat{g}_{2i} \hat{g}_{2i}' + \frac{1}{n} \sum_{i=1}^n (\rho_2(\bar{\lambda}'\hat{g}_{2i}) + 1) \hat{g}_{2i} \hat{g}_{2i}',$$

where by CS,

$$\frac{1}{n} \sum_{i=1}^n (\rho_2(\bar{\lambda}'\hat{g}_{2i}) + 1) \hat{g}_{2i} \hat{g}_{2i}' \leq \Big|\Big| \frac{1}{n} \sum_{i=1}^n \hat{g}_{2i} \hat{g}_{2i}' \Big|\Big| o_p(1).$$

By CS, M and Assumption 4.3(i),

$$\Big|\Big| \sum_{i=1}^n \hat{g}_{2i} \hat{g}_{2i}'/n \Big|\Big| \leq \frac{1}{n} \sum_{i=1}^n d_2(z_i)^2 \leq O_p(1).$$

Hence,

$$0 = -\frac{1}{n} \sum_{i=1}^n \hat{g}_{2i} + \Big(\frac{1}{n} \sum_{i=1}^n \hat{g}_{2i} \hat{g}_{2i}' + o_p(1) \Big) \hat{\lambda} \tag{B.3}$$

Now,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(\hat{g}_{2i}\hat{g}_{2i}' - g_{2i}g_{2i}') &= \frac{1}{n}\sum_{i=1}^{n} g_{2i}(\hat{g}_{2i}-g_{2i})' + \frac{1}{n}\sum_{i=1}^{n}(\hat{g}_{2i}-g_{2i})g_{2i}' + \frac{1}{n}\sum_{i=1}^{n}(\hat{g}_{2i}-g_{2i})(\hat{g}_{2i}-g_{2i})' \\ &:= E1 + E2 + E3.\end{aligned}$$

For $E1$,

$$\begin{aligned}||E1|| &\leq \frac{1}{n}\sum_{i=1}^{n}||g_{2i}|| \times ||\hat{g}_{2i}-g_{2i}|| \\ &\leq \Big(\frac{1}{n}\sum_{i=1}^{n} d_2(z_i)b_2(z_i)\Big)\big(||\hat{\theta}_2-\theta_{20}|| + ||\hat{h}_2-h_{20}||\big) \\ &= \big(\mathbb{E}[d_2(z)b_2(z)] + o_p(1)\big)\big(||\hat{\theta}_2-\theta_{20}|| + ||\hat{h}_2-h_{20}||\big) \\ &\leq o_p(1),\end{aligned}$$

where the first inequality follows by CS, the second by Assumptions 4.3(i), (iii), the equality by WLLN, and the third inequality by Assumption 4.4(i) and consistency of $\hat{\theta}_2$ and $\hat{h}_2$ for θ_{20} and h_{20}.

By identical arguments, $||E2|| \leq o_p(1)$.

For $E3$,

$$\begin{aligned}||E3|| &\leq \frac{1}{n}\sum_{i=1}^{n}||\hat{g}_{2i}-g_{2i}||^2 \\ &\leq \Big(\frac{1}{n}\sum_{i=1}^{n} b_2(z_i)^2\Big)\big(||\hat{\theta}_2-\theta_{20}|| + ||\hat{h}_2-h_{20}||\big)^2 \\ &= \big(\mathbb{E}[b_2(z)^2] + o_p(1)\big)\big(||\hat{\theta}_2-\theta_{20}|| + ||\hat{h}_2-h_{20}||\big)^2 \\ &\leq o_p(1),\end{aligned}$$

where the first inequality follows by CS, the second by Assumption 4.3(iii), the equality by WLLN, and the third inequality by Assumption 4.3(iii) and consistency of $\hat{\theta}_2$ and $\hat{h}_2$ for θ_{20} and h_{20}.

Hence, by CS,

$$\Big|\Big|\frac{1}{n}\sum_{i=1}^{n}(\hat{g}_{2i}\hat{g}_{2i}' - g_{2i}g_{2i}')\Big|\Big| \leq o_p(1).$$

By CS, WLLN, $(B.3)$ and Assumption 4.1(iii),

$$\hat{\lambda} = -\big(\Omega_2^{-1} + o_p(1)\big)\frac{1}{n}\sum_{i=1}^{n}\hat{g}_{2i}. \tag{B.4}$$

By a Taylor expansion around $\hat{\theta}_2 = \theta_{20}$,

$$\frac{1}{n}\sum_{i=1}^{n}\hat{g}_{2i} = \frac{1}{n}\sum_{i=1}^{n}\hat{g}_{2i}^{h} + \Big(\frac{1}{n}\sum_{i=1}^{n}G_{2i}(\bar{\theta}_2,\hat{h}_2)\Big)(\hat{\theta}_2 - \theta_{20}).$$

Note that

$$\begin{aligned}\Big|\Big|\Big(\frac{1}{n}\sum_{i=1}^{n}G_{2i}(\bar{\theta}_2,\hat{h}_2) - G_{2i}\Big)(\hat{\theta}_2-\theta_{20})\Big|\Big| &\leq \Big(\frac{1}{n}\sum_{i=1}^{n}\tilde{b}_2(z_i)\Big)\big(||\bar{\theta}_2-\theta_{20}|| + ||\hat{h}_2 - h_{20}||\big)||\hat{\theta}_2-\theta_{20}|| \\ &= \big(\mathbb{E}[\tilde{b}_2(z)] + o_p(1)\big)\big(||\bar{\theta}_2-\theta_{20}|| + ||\hat{h}_2 - h_{20}||\big)||\hat{\theta}_2-\theta_{20}|| \\ &\leq o_p(1)\end{aligned}$$

where the first inequality follows by Assumption 4.3(iv) and CS, the equality by WLLN, and the second inequality follows by Assumption 4.3(iv) and consistency of $\hat{\theta}_2$ (and hence $\bar{\theta}_2$), and $\hat{h}_2$.

Hence, by CS and WLLN,

$$\frac{1}{n}\sum_{i=1}^{n}\hat{g}_{2i} = \frac{1}{n}\sum_{i=1}^{n}\hat{g}_{2i}^{h} + [G_2 + o_p(1)](\hat{\theta}_2 - \theta_{20}).$$

Substituting this into $(B.4)$,

$$\hat{\lambda} = -(\Omega_2^{-1} + o_p(1))\Big(\frac{1}{n}\sum_{i=1}^{n}\hat{g}_{2i}^{h} + [G_2 + o_p(1)](\hat{\theta}_2 - \theta_{20})\Big). \tag{B.5}$$

Recall,

$$\rho_1(\hat{\lambda}'\hat{g}_{2i}) = -1 + \rho_2(\bar{\lambda}'\hat{g}_{2i})\hat{\lambda}'\hat{g}_{2i},\ (i = 1,...,n).$$

Therefore, by $(B.2)$,

$$\begin{aligned}0 &= -\frac{1}{n}\sum_{i=1}^{n}[-1 + \rho_2(\bar{\lambda}'\hat{g}_{2i})(\hat{\lambda}'\hat{g}_{2i})]\hat{G}_{2i}'\hat{\lambda} \\ &= -\frac{1}{n}\sum_{i=1}^{n}\hat{G}_{2i}'\hat{\lambda} + \Big(\frac{1}{n}\sum_{i=1}^{n}\rho_2(\bar{\lambda}'\hat{g}_{2i})\hat{G}_{2i}'\hat{g}_{2i}\Big)(\hat{\lambda}'\hat{\lambda}) \\ &= -\frac{1}{n}\sum_{i=1}^{n}\hat{G}_{2i}'\hat{\lambda} + \Big(\frac{1}{n}\sum_{i=1}^{n}\big(\rho_2(\bar{\lambda}'\hat{g}_{2i}) + 1\big)\hat{G}_{2i}'\hat{g}_{2i}\Big)(\hat{\lambda}'\hat{\lambda}) - \Big(\frac{1}{n}\sum_{i=1}^{n}\hat{G}_{2i}'\hat{g}_{2i}\Big)(\hat{\lambda}'\hat{\lambda}).\end{aligned}$$

Note that

$$\begin{aligned}\Big|\Big|\frac{1}{n}\sum_{i=1}^{n}\big(\rho_2(\bar{\lambda}'\hat{g}_{2i})+1\big)\hat{G}_{2i}'\hat{g}_{2i}\Big|\Big| &\leq \Big|\Big|\frac{1}{n}\sum_{i=1}^{n}\hat{G}_{2i}'\hat{g}_{2i}\Big|\Big|o_p(1)\\ &\leq \Big(\frac{1}{n}\sum_{i=1}^{n}d_2(z_i)\tilde{d}_2(z_i)\Big)o_p(1)\\ &= \big(\mathbb{E}[d_2(z)\tilde{d}_2(z)]+o_p(1)\big)o_p(1)\\ &\leq o_p(1),\end{aligned}$$

where the first inequality follows by CS, the second by Assumptions 4.3(i), (ii), the equality by WLLN, and the third inequality by Assumption 4.4(i).

Furthermore, by the above arguments,

$$\Big|\Big|\frac{1}{n}\sum_{i=1}^{n}\hat{G}_{2i}'\hat{g}_{2i}\Big|\Big| = O_p(1).$$

Hence,

$$0 = -\frac{1}{n}\sum_{i=1}^{n}\hat{G}_{2i}'\hat{\lambda}+O_p(||\hat{\lambda}||^2). \tag{B.6}$$

Now,

$$\begin{aligned}\Big|\Big|\frac{1}{n}\sum_{i=1}^{n}(\hat{G}_{2i}-G_{2i})'\hat{\lambda}\Big|\Big| &\leq \Big(\frac{1}{n}\sum_{i=1}^{n}\tilde{b}_2(z_i)\Big)\big(||\hat{\theta}_2-\theta_{20}||+||\hat{h}_2-h_{20}||\big)\times||\hat{\lambda}||\\ &= \big(\mathbb{E}[\tilde{b}_2(z)]+o_p(1)\big)\big(||\hat{\theta}_2-\theta_{20}||+||\hat{h}_2-h_{20}||\big)\times||\hat{\lambda}||\\ &\leq o_p(1),\end{aligned}$$

where the first inequality follows by CS and Assumption 4.3(iv), the equality by WLLN, and the second inequality by Assumption 4.3(iv), consistency of $\hat{\theta}_2$ and $\hat{h}_2$ for θ_{20} and h_{20}, and since $||\hat{\lambda}|| \leq n^{-\frac{1}{\alpha}}$ for some $\alpha > 2$ by definition of the search set Λ_n.

Hence, by CS and WLLN,

$$\frac{1}{n}\sum_{i=1}^{n}\hat{G}_{2i}'\hat{\lambda} = (G_2+o_p(1))'\hat{\lambda}$$

Substituting into $(B.6)$,

$$0 = -(G_2+o_p(1))'\hat{\lambda}+O_p(||\hat{\lambda}||^2). \tag{B.7}$$

By Lemma B.2, $\hat{\lambda} = O_p(n^{-\frac{1}{2}})$, so that $||\hat{\lambda}||^2 \leq o_p(n^{-\frac{1}{2}})$. Using this, and substituting $(B.5)$

into $(B.7)$,

$$\begin{aligned} o_p(n^{\frac{1}{2}}) &= (G_2 + o_p(1))'(\Omega_2^{-1} + o_p(1))\Big(\frac{1}{n}\sum_{i=1}^{n}\hat{g}_{2i}^h + [G_2 + o_p(1)](\hat{\theta}_2 - \theta_{20})\Big) \\ &= G_2'\Omega_2^{-1}\Big(\frac{1}{n}\sum_{i=1}^{n}\hat{g}_{2i}^h + G_2(\hat{\theta}_2 - \theta_{20})\Big), \end{aligned}$$

since $\sum_{i=1}^{n}\hat{g}_{2i}^h = O_p(n^{-\frac{1}{2}})$ by Lemma 5.1 of Newey [25].

Then, by Assumptions 4.1(ii), (iii),

$$\sqrt{n}(\hat{\theta}_2 - \theta_{20}) = -(G_2'\Omega_2^{-1}G_2)^{-1}G_2'\Omega_2^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{g}_{2i}^h + o_p(1). \tag{B.8}$$

Substituting $(B.8)$ into $(B.5)$,

$$\begin{aligned} \sqrt{n}\hat{\lambda} &= -(\Omega_2^{-1} + o_p(1))\Big(\frac{1}{n}\sum_{i=1}^{n}\hat{g}_{2i}^h - [G_2 + o_p(1)]\Big)\Big\{(G_2'\Omega_2^{-1}G_2)^{-1}G_2'\Omega_2^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{g}_{2i}^h + o_p(1)\Big\} \\ &= -[\Omega_2^{-1} - \Omega_2^{-1}G_2(G_2'\Omega_2^{-1}G_2)^{-1}G_2'\Omega_2^{-1}]\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{g}_{2i}^h + o_p(1) \\ &= -P\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{g}_{2i}^h + o_p(1). \end{aligned}$$

□

Lemma B.6 (***GEL Implied Probabilities (Ramalho and Smith [33], Lemma A.1)***). *Under Assumptions 4.1-4.4, the GEL Implied Probabilities satisfy*

$$\hat{\pi}_i = \frac{1}{n} + \frac{1}{n}\hat{g}_{2i}'\hat{\lambda}(1 + o_p(1)) + O_p(n^{-2}) \ \ (i = 1, ..., n).$$

Proof. The derivation is identical to Lemma A.1. of Ramalho and Smith [33] using Lemma B.1, $||\sum_{i=1}^{n}\hat{g}_{2i}/n|| = O_p(n^{-\frac{1}{2}})$ from Lemma B.3 and $\hat{\lambda} = O_p(n^{-\frac{1}{2}})$ by Lemma B.2. □

Ph.D. Thesis: Chapter Three

Copula-Graphic Inference with Cause-specific Hazard Models

Ashish Patel[1*] Chien-Ju Lin[1,2] James Wason[1,2,3]

University of Cambridge[1]
MRC Biostatistics Unit[2]
University of Newcastle[3]

Abstract

Time-to-event studies with several possible causes of event for each subject describe many problems in research. When these competing risks are dependent and only information on the time-to-first-event is available, marginal survival functions cannot be identified (Cox [7], Tsiatis [22]). Copula-Graphic estimators (Zheng and Klein [24]) exploit information on the dependence structure between risks to return consistent estimators. This paper derives asymptotic results for a class of parametric Copula-Graphic estimators and considers the use of confidence intervals for inference on conditional marginal survival functions. The efficacy of the asymptotic confidence intervals and their sensitivity to the choice of copula parameter are illustrated by simulation.

Keywords: Competing Risks, Copula-Graphic Estimators, Marginal Survival Analysis, Cause-specific Hazard Functions

[*]This work represents my contribution to a larger project on improving phase II oncology trials by efficient use of continuous tumour size data, conducted by James Wason and Chien-Ju Lin; I thank them for their guidance on this topic. I thank Debopam Bhattacharya for helpful discussions and Richard Smith for detailed comments.

1 Introduction

Consider time-to-event studies where there are several possible causes for the event. The possible causes for the event are described as competing risks; the occurrence of one cause of an event precludes other causes of the event from being observed. Therefore, only the information on the time to the first cause of event is available. Usually, this information consists of data on i) time to the first cause of event, and ii) which cause of event occurs. In such a situation, researchers are often concerned with inference on a particular cause of event. The central concern of this paper is the estimation of conditional marginal survival functions under dependent competing risks. Such functions are of great interest, for example, in biostatistics and actuarial science; analysis on the cause of death is important for informing future treatment or pricing policies. In economics, models of competing risks have been popular for studies of unemployment duration.

When competing risks are independent, standard approaches can be extended for estimation and inference of marginal survival functions. However, in many contexts the assumption that the competing risks are independent is not plausible. For example, Lo et al. [16] study the effects of German labour market reforms on latent marginal distributions of competing risks to leave unemployment. The competing risks to leave unemployment are 'recall to previous employer', 'low-wage full-time employment', 'self-employment', and others. They argue two or more of these risks are likely to be related.

Inference in a dependent competing risks context is more problematic than when risks are assumed to be independent. The non-identification theorem, see Cox [7] and Tsiatis [22], states that for any joint distribution of latent event times, there exists another such distribution of *independent* latent event times that yields the same distribution of the observable data (namely, time to event and which event occurs). This implies the relationship between latent event times cannot be identified using only observed data. Crucially this constitutes a missing data problem which is insurmountable, without the imposition of further untestable restrictions, for consistent estimation of marginal survival functions of individual causes. However, with some knowledge of the dependence structure between competing risks, then marginal survival functions for specific causes can be recovered.

The copula is a nonparametric function which captures all information about the dependence between two random variables. This paper considers copula families that are parameterised by a finite-dimensional parameter, henceforth, the copula parameter. In a novel approach, assuming knowledge of the copula of competing risks, the class of Copula-Graphic (CG) estimators (Zheng and Klein [24]) consistently estimates marginal survival functions based on observable quantities. CG methods can also be used to handle dependent censoring by characterising censoring as an individual event. Many extensions of the CG estimator have been studied. For example, Rivest and Wells [21] derive asymptotic properties of the CG estimator with nonparametrically-estimated cumulative incidence functions (CIFs, see Section 2.2) for

the class of Archimedean copulas. Chen [6] considers a semiparametric model for the cumulative hazard function. Liu and Wang [15] study a setting in which event times may be missing. Braekers and Veraverbeke [3] derive results for kernel-based estimation that incorporates covariate information. In closely related work, Lo and Wilke [17] extend CG estimators to deal with more than two dependent risks and covariates in a regression setting.

Under CG methods, functionals of the CIF are nonlinearly combined by an assumed copula. The CIF is identified by observable data and can therefore be estimated nonparametrically as in Zheng and Klein [24]. However, if there are many explanatory variables to consider or if the sample size is not sufficient then in practice parametric assumptions are likely to be imposed. This paper considers a parametric approach to modelling the CIF where cause-specific hazard functions are modelled (cf. Jeong and Fine [12]).

The copula plays a central role in shaping conditional marginal survival function estimates in CG methods. Yet most of the developments in CG methods have not relaxed the requirement that the copula function be fully known. To the best of our knowledge, only one attempt to incorporate uncertainty in copula functions has been made; Chaieb et al. [5] estimates the copula parameter via an estimating equation obtained by equating two equivalent expressions for Kendall's tau. However, Chaieb et al.'s [5] approach is restricted to Archimedean copulas with one-dimensional parameters, and their framework does not allow for covariates. Such copula specifications may not be sufficiently general to model more complicated dependence structures that may be empirically relevant. This may be a problem for marginal analysis of transitions to employment. In Lo et al.'s [16] study of German labour market reforms discussed above, transitions to low-wage employment may depend less on high-wage full-employment options initially, but under an extended spell of unemployment duration, the lack of high-wage options may expedite transitions to low-wage employment. For such studies, researchers may consider CG estimation with the Joe-Clayton copula which has two parameters permitting asymmetric behaviour in the upper and lower tails.

Therefore it is important to examine how robust CG estimators are to the choice of the copula parameter. This paper studies CG methods for conditional marginal survival function estimation for Archimedean copulas. We derive asymptotic confidence intervals for a class of parametric Copula-Graphic estimators and consider its efficacy for inference on conditional marginal survival functions. The performance of the proposed method is evaluated by average coverage in simulation results, with particular focus on the sensitivity of the results to choices of the copula parameter. The methods and results obtained in this paper contribute to the understanding of marginal survival function estimation in dependent competing risks models, and complement the results of Lo and Wilke [17] by considering inference based on confidence intervals.

The paper is organised as follows. Section 2 describes the dependent competing risks model and CG estimation. Section 3 discusses estimation of cause-specific hazard and conditional

marginal survival functions and presents asymptotic results. Section 4 illustrates the use of the proposed method by simulation. Section 5 concludes. All proofs are given in the Appendix.

The following abbreviations are used. $\xrightarrow{p}$: converges in probability to, $\xrightarrow{d}$: converges in distribution to, T : the triangle inequality, CS: the Cauchy-Schwarz inequality, UWL: the uniform weak law of large numbers (for example, Lemma 2.4 of Newey and McFadden [19]), WLLN: the weak law of large numbers, CMT: the continuous mapping theorem, LHS: left hand side, RHS: right hand side, and $||.||$ is the Euclidean norm. For a vector v, $vv' = v^{\otimes 2}$. For a d_λ-dimensional vector $\lambda \in \Lambda$ and a d_ψ-dimensional vector of functions $\psi : \Lambda \to \mathbb{R}^{d_\psi}$, $\nabla_\lambda \psi(\lambda)$ denotes the $d_\psi \times d_\lambda$ matrix of derivatives $\partial\psi(\lambda)/\partial\lambda$.

2 Competing Risks Model

Consider identification and estimation of a dependent, competing risks model. In particular, there are several possible causes for an event to occur. For notational convenience, only the case of two causes (or risks) are considered, $j \in \{1, 2\}$. This can easily be extended to a model with J causes (see, for example, Lo and Wilke [17]) at no cost to the efficiency properties derived here. Let $T^{(j)}$ denote the latent time to event from cause j. The researcher does not have access to all data on $T^{(j)}$ $(j = 1, 2)$. Instead, the only data available are the time to first event, $T = \min\{T^{(1)}, T^{(2)}\}$, whichever event happens first, i.e. $\delta = \arg\min_{j\in\{1,2\}}\{T^{(j)}\}$, and a d_X-dimensional vector of covariates, $X \in \mathcal{X}$, for n individuals. Survival functions can be estimated by standard methods when $T^{(1)}$ and $T^{(2)}$ are independent of each other, that is, when $S(t_1, t_2) = S_1(t_1)S_2(t_2)$ is satisfied, where $S(t_1, t_2) = \mathbb{P}(T^{(1)} > t_1, T^{(2)} > t_2)$ is the joint survival function, and $S_j(t_j) = \mathbb{P}(T^{(j)} > t_j)$ $(j = 1, 2)$ are the marginal survival functions. When $T^{(1)}$ and $T^{(2)}$ are correlated, marginal survival functions cannot be identified without further restrictions. CG estimators (Zheng and Klein [24]) exploit knowledge of the dependence structure between the competing risks to return consistent estimators of marginal survival functions.

2.1 Archimedean copulas

Consider the conditional marginal survival functions, $S_j(t, x) = \mathbb{P}(T^{(j)} > t \,|X = x)$ $(j = 1, 2)$, and the conditional joint survival function $S(t, x) = \mathbb{P}(T > t \,|X = x)$.

Under regularity conditions, if the true marginal distributions are continuous, there exists a unique copula $\mathcal{C} : [0, 1]^2 \times \mathcal{X} \to [0, 1]$ such that

$$S(t, x) \quad = \quad \mathcal{C}(S_1(t, x), S_2(t, x); x), \tag{2.1}$$

see, for example, Theorem 1 of Embrechts ([8], p.641). Copulas are a distribution function on the unit square, taking and returning values in the interval $[0, 1]$. For the purpose of deriving

closed-form solutions and asymptotic properties of CG estimators, many previous works have focused on the class of copula functions that belong to the Archimedean family. The following assumption is as stated in Lo and Wilke [17].

Assumption 2.1. *The copula function $\mathcal{C}_\alpha(S_1(t,x), S_2(t,x);\, x)$ that satisfies (2.1) is indexed by a d_α-dimensional vector parameter α and belongs to the family of Archimedean copulas. Furthermore, the copula does not depend directly on x, but only indirectly through S_1 and S_2. That is,*

$$\mathcal{C}_\alpha(S_1(t,x), S_2(t,x);\, x) = \phi_\alpha^{-1}[\phi_\alpha(S_1(t,x)) + \phi_\alpha(S_2(t,x))] \tag{2.2}$$

for $\phi_\alpha : [0,1] \times \mathcal{A} \to [0,1]$, where $\mathcal{A}$ is a compact subspace of $\mathbb{R}^{d_\alpha}$, and ϕ_α is strictly decreasing and a twice differentiable, continuous function such that $\phi_\alpha(1) = 0$.

The function $\phi_\alpha(u) = \phi(u;\alpha)$ is known as the generating function of the copula. Assumption 2.1 contains many widely-used copulas used in practice. The assumption that the copula function does not depend directly on x is only maintained to facilitate simpler analysis of marginal effects, as in Lo and Wilke [17]. However, this can be relaxed at the cost of extra derivations and notation.

Example 2.1. **(i)** Frank copula: $\phi_\alpha(t) = -\ln\left[\frac{\exp(-\alpha t)-1}{\exp(-\alpha)-1}\right]$ for $\alpha \in (-\infty,\infty)\backslash 0$, and $\phi_\alpha(t) = -\ln(t)$ for $\alpha = 0$. Although risks are allowed to be positively or negatively linked, no dependence is allowed in the tails; **(ii)** Clayton copula: $\phi_\alpha(t) = (t^\alpha - 1)\alpha$ for $\alpha > 0$. Risks are only allowed to be positively linked, but dependence in the lower tail is permitted.

2.2 Cause-specific hazard functions

A key ingredient in CG methods is the cumulative incidence function (CIF). The CIF of cause j is the probability that the event occurs due to cause j by a given time t conditional on covariates x, $Q_j(t,x) = \mathbb{P}(T \le t, \delta = j \,|X = x)$, $j \in \{1,2\}$. CIF curves thus show the cumulative cause-specific event rates over time. Functionals of CIFs are often of independent interest, for example for assessing covariate effects on the competing risks. There are several ways of modelling the CIF, including direct specifications (see, for example, Jeong and Fine [13]), and through modelling the subdistribution hazard functions (Fine and Gray [9]). See Zhang et al. [23] for a review of existing methods.

The cause-specific hazard function measures the rate at which a particular cause of event occurs when other competing risks can also occur. For cause j it is defined as

$$h_j(t,x) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \le T \le t + \Delta t,\, \delta = j \,\big|\, T \ge t, X = x)}{\Delta t}\ ,\ j \in \{1,2\}.$$

This hazard function describes the instantaneous probability that an event of type j occurs at time t for an individual with characteristics x. These quantities are identified from competing

risk data. A standard approach is to specify regression models for the cause-specific hazard functions, which leads to plug-in estimation of the CIF. This is the approach followed here, although other approaches can be used to construct an analogous estimator to the one proposed in this paper.

Consider a parametric specification for the cause-specific hazard functions (cf. Jeong and Fine [12]), i.e., for cause $j \in \{1, 2\}$,

$$h_j(t, x) := h_j(t, x; \theta_{j0})$$

where h_j is continuously differentiable in $\theta_{j0} \in \Theta_j$, where $\dim(\theta_j) = d_{\theta_j}$ and Θ_j is a compact subspace of $\mathbb{R}^{d_{\theta_j}}$. Define the pseudo-marginal survival function for cause j, $j = \{1, 2\}$, $W_j(t, x; \theta_{j0}) = \exp\Big(- \int_0^t h_j(u, x; \theta_{j0}) du \Big)$, which leads to the following representation of the conditional joint survival function, $S(t, x; \theta_0) = \mathbb{P}(T > t | \theta_0; X = x)$,

$$S(t, x; \theta_0) = \prod_{j=1}^{2} W_j(t, x; \theta_{j0}) \tag{2.3}$$

where $\theta_0 = (\theta_{10}', \theta_{20}')' \in \Theta = \Theta^{d_{\theta_1}} \times \Theta^{d_{\theta_2}}$. The CIF for cause j is given by

$$\begin{aligned} Q_j(t, x; \theta_0) &= \int_0^t h_j(u, x; \theta_{j0}) S(u, x; \theta_0) du \\ &= \int_0^t h_j(u, x; \theta_{j0}) \prod_{k=1}^{2} W_k(u, x; \theta_{k0}) du. \end{aligned} \tag{2.4}$$

Without loss of generality, the focus here is on the conditional marginal survival function for cause 1 conditional on $X = x$ given by

$$S_1(t, x; \theta_0, \alpha) = \phi_\alpha^{-1}\Big[- \int_0^t \phi_\alpha'(S(u, x; \theta_0)) S(u, x; \theta_0) h_1(u, x; \theta_{10}) du \Big], \tag{2.5}$$

where $\phi_\alpha'(u) = \partial \phi_\alpha(u) / \partial u$. For derivation of (2.5) see Lo and Wilke ([17], p.39).

Under Assumption 2.1, CG methods combine observable quantities through copula functionals to return consistent estimates of conditional marginal survival functions.

2.3 Example: Proportional odds model with Weibull baseline hazard

Consider the class of parametric models generated by the proportional odds assumption coupled with a Weibull baseline hazard model. The proportional odds model is particularly popular for ordinal response variables, see, for example, Ananth and Kleinbaum [1]. The flexibility of the Weibull specification has proved popular for modelling unemployment durations, see, for example, Cameron and Trivedi ([4], pp.603-608). The baseline hazards from the Weibull distribution are $h_{oj}(t) = \rho_{j0} v_{j0} t^{\rho_{j0} - 1}$ $(j = 1, 2)$. Hence, following Section 3.1.1 of Lo and

Wilke [17], the hazard functions are given by $h_j(t,x;\theta_{j0}) = h_{0j}(t)\exp(x'\beta_{j0})W_j(t,x;\theta_{j0})$ and the pseudo-marginal survival function is $W_j(t,x;\theta_{j0}) = \{1+\exp(x'\beta_{j0})v_j t^{\rho_{j0}}\}^{-1}$ $(j=1,2)$. The conditional joint survival function is then $S(t,x;\theta_0) = W_1(t,x;\theta_{10})W_2(t,x;\theta_{20}) = \{1+\exp(x'\beta_{10})v_{10}t^{\rho_{10}}\}^{-1}\{1+\exp(x'\beta_{20})v_{20}t^{\rho_{20}}\}^{-1}$.

For the dependence between risks, consider the Clayton copula which allows for extreme lower tail correlations, see Example 2.1(ii). The relevant functionals of the Clayton copula for recovering the conditional marginal survival functions are $\phi_\alpha^{-1}(u) = (\alpha u+1)^{-\frac{1}{\alpha}}$ and $\phi'_\alpha(u) = -u^{-(\alpha+1)}$. Since the scalar copula parameter α is assumed known, the unknown parameter vector is then the $(d_{\theta_1}+d_{\theta_2}+4)$-dimensional vector $\theta_0 = (\theta'_{10},\theta'_{20})$ where $\theta_{j0} = (\beta'_{j0}, v_{j0}, \rho_{j0})'$ $(j=1,2)$.

From (2.4), the conditional marginal survival function for cause 1 is

$$\begin{aligned} S_1(t,x;\theta_0,\alpha) &= \left(\alpha\int_0^t \Big(\{1+\exp(x'\beta_{10})v_1u^{\rho_{10}}\}^{-1}\{1+\exp(x'\beta_{20})v_2u^{\rho_{20}}\}^{-1}\Big]^{-(\alpha+2)}\rho_{10}\right. \\ &\quad \left.\times v_{10}t^{\rho_{10}-1}\exp(x'\beta_{10})\{1+\exp(x'\beta_{10})v_{10}t^{\rho_{10}}\}^{-1}\Big)du+1\right)^{-\frac{1}{\alpha}}. \end{aligned}$$

A consistent estimator for $S_1(t,x;\theta_0,\alpha)$ is obtained by substitution of the estimate of θ_0 described in the following section.

3 Copula-Graphic Confidence Intervals

Since θ_{j0}, $(j=1,2)$, are independently identified from the cause-specific hazards $h_j(t,x)$, and the conditional marginal survival function is a function of θ_0, θ_0 is first estimated by standard procedures. Given an estimator $\hat\theta$ of θ_0 and (2.5), an estimator of the conditional marginal survival function $S_j(t,x;\theta_0,\alpha)$ is $S_j(t,x;\hat\theta,\alpha)$. This section derives the asymptotic distribution of $S_j(t,x;\hat\theta,\alpha)$ which is then used to construct confidence intervals for the conditional marginal survival function $S_j(t,x;\theta_0,\alpha)$.

3.1 Estimation of θ_0

Cause-specific hazard functions can be estimated by maximum likelihood. Given an i.i.d. sample of n individuals, the observable data is $\{t_i,\delta_i,x_i\}_{i=1}^n$, where for each individual $(i=1,...,n)$, $T=t_i$ is the observed time to event, $\delta_i = \arg\min_j\{T^{(j)}\}$ reveals the cause of event, and x_i is a d_X-vector of covariates. Let $\epsilon_{1i} = \mathbb{I}\{\delta_i=1\}$ and $\epsilon_{2i} = \mathbb{I}\{\delta_i=2\}$, $i=1,...,n$; thus $\epsilon_{2i} = 1-\epsilon_{1i}$.

For $\theta=(\theta_1,\theta_2)\in\Theta$, as in Jeong and Fine ([12], p.191-2), the full likelihood function can be

written

$$L_n(\theta) = \prod_{i=1}^{n} \prod_{j=1}^{2} h_j(t_i, x_i; \theta_j)^{\epsilon_{ji}} W_j(t_i, x_i; \theta_j). \tag{3.1}$$

Since the score equation for θ_1 does not depend on θ_2 and vice versa, estimation of the cause-specific hazards can be done separately. For $j = \{1, 2\}$, an estimate $\hat{\theta}_j$ of θ_{j0} satisfies the score equations,

$$U_{jn}(\hat{\theta}_j) = 0$$

where $U_{jn}(\theta_j) = \nabla_{\theta_j} \log L_n(\theta_j) = \sum_{i=1}^{n} \epsilon_{ji} \Big(\frac{\nabla_{\theta_j} h_j(t_i, x_i; \theta_j)}{h_j(t_i, x_i; \theta_j)} \Big) - \sum_{i=1}^{n} \int_0^{t_i} \nabla_{\theta_j} h_j(u, x_i; \theta_j) du$.

3.2 Model assumptions

Let $\mathbb{E}[.]$ denote the expectation taken with respect to X. It is assumed that $\tau \in \mathbb{R}_+$ satisfies $T \leq \tau$ for all possible event times $T > 0$. Let $N_{ji}(t)$ be the right-continuous process that indicates whether the i-th individual observes an event of cause j by time t, $N_{ji}(t) = \mathbb{I}(T_i \leq t, \delta_i = j)$. Let $Y_i(t)$ be the left-continuous at-risk process that indicates whether the i-th individual is at risk of the event at time t, see Kalbfleisch and Prentice ([14], Chapter 8, Section 8.2.7, p.265).

Assumption 3.1. *For $j = \{1, 2\}$,* **(i)** *$\{t_i, \delta_i, x_i\}_{i=1}^n$ is an i.i.d. sample such that $\mathbb{P}(t_i > t|x) = \exp(-\int_0^t \sum_{j=1}^J h_j(u, x; \theta_{j0}) du)$;* **(ii)** *for any $t \in [0, \tau]$ and $x \in \mathcal{X}$, if $\theta_j \neq \theta_{j0}$ then $h_j(t, x; \theta_j) \neq h_j(t, x; \theta_{j0})$;* **(iii)** *$\theta_{j0} \in int(\Theta_j)$ where Θ_j is a compact subspace of $\mathbb{R}^{d_{\theta_j}}$;* **(iv)** *$h_j(t, x; \theta_j)$ is continuous at each $\theta_j \in \Theta_j$ with probability one and $|h_j(t, x; \theta_j)| \leq d_0(x)$ and $\mathbb{E}[d_0(x)] < \infty$, uniformly $t \in [0, \tau]$;* **(v)** *$h_j(t, x; \theta_j) > 0$ for all $x \in \mathcal{X}$, $\theta \in \Theta$, uniformly $t \in [0, \tau]$;* **(vi)** *$h_j(t, x; \theta_j)$ is thrice differentiable in a neighbourhood $\mathcal{N}$ of the true value θ_0,* **(vii)** *condition (iv) holds with $\nabla_{\theta_j} h_j(t, x; \theta_j)$ replacing $h_j(t, x; \theta_j)$ and $d_1(x)$ replacing $d_0(x)$ such that $||\nabla_{\theta_j} h_j(t, x; \theta_j)|| \leq d_1(x)$ and $\mathbb{E}[d_1(x)] < \infty$;* **(viii)** *condition (iv) holds with $\nabla_{\theta_j \theta_j'} h_j(t, x; \theta_j)$ replacing $h_j(t, x; \theta_j)$ and $d_2(x)$ replacing $d_0(x)$ such that $||\nabla_{\theta_j \theta_j'} h_j(t, x; \theta_j)|| \leq d_2(x)$ and $\mathbb{E}[d_2(x)] < \infty$;* **(ix)** *$\Sigma_j(\theta_j, \tau) = \int_0^\tau \mathbb{E}\big[[\nabla_{\theta_j \theta_j'} h_j(u; x, \theta_j)/h_j(u; x, \theta_j)] Y(u) \{h_j(u; x, \theta_j) - h_j(u; x, \theta_{j0})\}\big] du + \int_0^\tau \mathbb{E}\big[(\nabla_{\theta_j} h_j(u; x, \theta_j)/h_j(u; x, \theta_j))^{\otimes 2} Y(u) h_j(u; x, \theta_{j0})\big] du$ is positive definite for all $\theta_j \in \Theta_j$;* **(x)** *differentiation with respect to θ and integration with respect to t can be interchanged for all $\theta \in \mathcal{N}$.*

Assumption 3.1(i)-(ix) provides a set of sufficient conditions for consistency of maximum likelihood estimators of cause-specific hazard functions, and contain additional assumptions which are required for asymptotic normality. The conditions relate to Assumption C1 of Kalbfleisch and Prentice ([14], p.175) and Theorems 2.5 and 3.1 of Newey and McFadden [19]. The conditions are also sufficient for uniform law of large numbers arguments required for derivations. Assumption 3.1(x) can be relaxed by verifying conditions required for the dominated convergence theorem and Fubini's theorem, see Section 5.8 of Kalbfleisch and Prentice ([14], p.179).

Assumption 3.1(iv), (vii), (viii) are continuity and boundedness conditions required for uniform law of large numbers arguments. Assumption 3.1 is fulfilled by many popular parametric hazard specifications.

Assumption 3.2. *Let $\nabla_{\theta_{jk}}$ be the k-th element of the parameter vector θ_j. For $j \in \{1,2\}$, for all $k \in \{1,...,d_{\theta_j}\}$ and $\epsilon > 0$,*

$$n^{-1}\sum_{i=1}^{n}\int_0^\tau \left(\nabla_{\theta_{jk}} \log h_j(u,x_i;\theta_{j0})\right)^2 \mathbb{I}\Big\{\Big|n^{-\frac{1}{2}}\nabla_{\theta_{jk}} \log h_j(u,x_i;\theta_{j0})\Big| > \epsilon\Big\} Y_i(u)h_j(u,x_i;\theta_{j0}) \xrightarrow{p} 0.$$

Assumption 3.2 corresponds to Condition 2 in Section 5.8 of Kalbfleisch and Prentice ([14], p.180) and is one of the conditions of Rebolledo's martingale central limit theorem (see Lemma A1, Section A.1 of the appendix). The process on the LHS of the limit above is similar to the predictable variation process of the score $U_{jn}(\theta_j,\tau)$ (see proof of Lemma 3.1(ii) in the appendix). However, the indicator function ensures the process only registers jumps of $U_{jn}(\theta_j,\tau)$ that are at least ϵ in size.

The condition guarantees the contribution of any single process is negligible in the limit. Therefore, it is a type of Lindeberg condition that allows a central limit theorem to hold for the score process evaluated at the true parameter θ_{j0}. Assumptions 3.1 and 3.2 are sufficient conditions for asymptotic normality of maximum likelihood estimators of cause-specific hazard functions.

Lemma 3.1 (Cause-specific hazard and CIF). *Under Assumptions 3.1 and 3.2, for $j \in \{1,2\}$, **(i)** $\hat{\theta}_j$ is consistent for θ_{j0}; **(ii)** $\sqrt{n}(\hat{\theta}_j - \theta_{j0}) \xrightarrow{d} \mathcal{N}(0, \Sigma_j(\theta_{j0})^{-1})$ where $\Sigma_j(\theta_{j0}) = \int_0^\tau \mathbb{E}\big[Y(u)\nabla_{\theta_j} h_j(u,x;\theta_{j0})^{\otimes 2}/h_j(u,x;\theta_{j0})\big]du$; **(iii)** $\sqrt{n}(Q_j(t,x;\hat{\theta}) - Q_j(t,x;\theta_0)) \xrightarrow{d} \mathcal{N}\big(0, [\nabla_\theta Q_j(t,x;\theta_0)]\Sigma(\theta_0)^{-1}[\nabla_\theta Q_j(t,x;\theta_0)]'\big)$, where $\Sigma(\theta_0) = diag(\Sigma_1(\theta_{10}), \Sigma_2(\theta_{20}))$, and $\Sigma_j(\theta_{j0}) = \Sigma_j(\theta_{j0},\tau)$ is defined in Assumption 3.1(ix).*

The third conclusion follows immediately from the second using the delta method. The proof of parts (i) and (ii) is outlined in Kalbfleisch and Prentice ([14], pp.172-7 and pp.179-180); however a full proof is provided in the Appendix for completeness, along with an explicit expression for $\nabla_\theta Q_j(t,x;\theta_0)$.

Note that the CIFs depend on both cause-specific hazard parameters θ_{10} and θ_{20}. This means that for consistent estimation of the CIF for any cause, correct specification of both cause-specific hazard functions is required. This requirement can be avoided by direct modelling of the CIF, however, this imposes ad-hoc restrictions on the parameter sets Θ_1 and Θ_2.

3.3 Conditional marginal survival functions

The estimator $\hat{\theta}$ of θ_0 can be plugged in to equation (2.5) to obtain an estimator $S_j(t, x; \hat{\theta}, \alpha)$ of the conditional marginal survival function $S_j(t, x; \theta_0, \alpha)$, $(j = 1, 2)$.

Corollary 3.1 (Conditional marginal survival function). *Under Assumptions 2.1, 3.1-3.2, given the estimator $\hat{\theta}$ of Section 3.1, the estimator $S_j(t, x; \hat{\theta}, \alpha)$ for $S_j(t, x; \theta_0, \alpha)$, $(j = 1, 2)$* ***(i)*** *is consistent for $S_j(t, x; \theta_0, \alpha)$ (2.5);* ***(ii)*** *is asymptotically normal such that*

$$\sqrt{n}(S_j(t, x; \hat{\theta}, \alpha) - S_j(t, x; \theta_0, \alpha)) \xrightarrow{d} \mathcal{N}(0, \mathcal{V}_{S_j}(t, x))$$

where $\mathcal{V}_{S_j}(t, x)$ is defined by equations $(A.11) - (A.13)$ in the proof of Corollary 3.1 in the Appendix.

Asymptotic confidence intervals can be constructed in the usual way. For an individual with covariate x_0, a $(1 - \gamma)$-confidence interval for the conditional marginal survival function $S_j(t, x_0; \theta_0, \alpha)$ $(j = 1, 2)$ is given by

$$\left[S_j(t, x_0; \hat{\theta}, \alpha) - z_{\frac{\gamma}{2}} \sqrt{\frac{\hat{\mathcal{V}}_{S_j}(t, x_0)}{n}}, S_j(t, x_0; \hat{\theta}, \alpha) + z_{\frac{\gamma}{2}} \sqrt{\frac{\hat{\mathcal{V}}_{S_j}(t, x_0)}{n}} \right],$$

where z_γ is the $(1 - \gamma)$-th quantile of the normal distribution, and $\hat{\mathcal{V}}_{S_j}(t, x)$ is a consistent estimator for $\mathcal{V}_{S_j}(t, x)$.

The choice of the copula function to describe the dependence between competing risks remains important. McNeil and Neslehova [18] provides some instruction on the shape constraints implied by various functions within the Archimedean family. Some crude bounds on marginal survival functions also constrain the range of copulas that are feasible. For example, for any consistent estimators $\hat{S}_j(t, x)$ of the conditional marginal survival functions $S_j(t, x)$ $(j = 1, 2)$, Frèchet-Hoeffding bounds imply, asymptotically, for any $t \in [0, \tau]$ and covariates $x_0 \in \mathcal{X}$,

$$\hat{S}_1(t, x_0) + \hat{S}_2(t, x_0) - 1 \leq \hat{S}(t, x_0) \leq \min\left\{\hat{S}_1(t, x_0), \hat{S}_2(t, x_0)\right\}$$

where $\hat{S}(t, x_0)$ is a nonparametric estimator of the conditional joint survival probability $\mathbb{P}(T^{(1)} > t, T^{(2)} > t \,|X = x_0)$. In practice, some known characteristics of the relationship between the competing risks may help govern copula choice.

3.4 Marginal effects

A key question in research often concerns how covariates impact on survival probabilities. For example, a policy issue may be how age affects the probability of finding a job. In medical

research, it is of key interest to see how prognosis depends on clinical variables. For the case where a covariate x_k, $k \in \{1, ..., d_X\}$ is continuous, a consistent estimator for the average marginal effect of a change in a covariate value on survival probabilities for cause j is

$$\frac{1}{n}\sum_{i=1}^{n} \nabla_{x_k} S_j(t, x_i; \hat{\theta}, \alpha),$$

where $\nabla_{x_k} S_j(t, x; \theta, \alpha)$ can be obtained as follows.

Let $A_j(t, x; \theta, \alpha) := -\int_0^t \phi'_\alpha(S(u, x; \theta)) Q'_j(u, x; \theta) du$ so that $S_j(t, x; \theta, \alpha) = \phi_\alpha^{-1}(A_j(t, x; \theta, \alpha))$. Let $\phi_\alpha^{-1(\prime)}(u) = \partial \phi_\alpha^{-1}(u)/\partial u$ and $\phi''_\alpha(u) = \partial^2 \phi_\alpha(u)/\partial u^2$. Then

$$\nabla_{x_k} S_j(t, x; \theta, \alpha) = \phi_\alpha^{-1(\prime)}(A_j(t, x; \theta, \alpha)) \big[\nabla_{x_k} A_j(t, x; \theta, \alpha)\big]$$

where

$$\begin{aligned} \nabla_{x_k} A_j(t, x; \theta, \alpha) \;\; = \;\; & -\int_0^t \phi''_\alpha(S(u, x; \theta)) \big[\nabla_{x_k} S(u, x; \theta)\big] Q'_j(u, x; \theta) du \\ & -\int_0^t \phi'_\alpha(S(u, x; \theta)) \big[\nabla_{x_k} Q'_j(u, x; \theta)\big] du. \end{aligned}$$

Since $S(u, x; \theta) = \exp\big(-\int_0^u \sum_{j=1}^2 h_j(s, x; \theta_j)\big)$ and $Q'_j(u, x; \theta) = S(u, x; \theta) h_j(u, x; \theta_j)$,

$$\nabla_{x_k} S(u, x; \theta) = -\Big(\int_0^u \sum_{j=1}^{2} \nabla_{x_k} h_j(u, x; \theta_j) du\Big) S(u, x; \theta)$$

and

$$\nabla_{x_k} Q'_j(u, x; \theta) = \big[\nabla_{x_k} S(u, x; \theta)\big] h_j(u, x; \theta_j) + S(u, x; \theta) \big[\nabla_{x_k} h_j(u, x; \theta_j)\big].$$

For the derivation of these expressions, see the results of Lo and Wilke ([17], Proposition 2, p.24).

4 Simulation study

The primary goal of this simulation study is to evaluate the performance of CG-estimated asymptotic confidence intervals, henceforth CG confidence intervals, as a tool for finite sample inference on the conditional marginal survival function. By applying our results from Section 3.3 to construct confidence intervals, the CG estimator of the conditional marginal survival function is compared with two nonparametric (kernel-based) conditional marginal survival function estimators[1]; one based on the observable competing risks data, and one based on the unobservable, full event times $\{T_i^{(1)}, T_i^{(2)}\}_{i=1}^n$.

[1]Kernel CDF estimation was computed using Hayfield and Racine's [11]'s *np* R package; the optimal bandwidth was computed by least squares cross validation.

Throughout this section, the results for 95%-confidence intervals are studied. The results of the simulation study show that the performance of CG confidence intervals may be highly dependent on the correlation structure implied by the copula, and highlight the difficulty of obtaining good coverage for the tails of the distribution.

The simulation study of Lo and Wilke ([17], p.34, Section 4) discusses the large biases of CG-estimated conditional marginal survival functions as the chosen copula parameter α moves away from its true value. Interestingly, they found that incorrect assumptions about the copula parameter value led to larger biases than incorrect assumptions on the copula family. The results here provide further evidence that the choice of copula parameter remains important, however, coverage under the correct choice of copula parameter is not uniformly higher at all regions of a conditional marginal survival function.

4.1 Copulas and the cause-specific hazard function

Covariate values are generated from a truncated normal distribution $X \sim N(0,1)$ such that $-2 \leq X \leq 2$. For the cause-specific hazard specification, we consider the Weibull distribution for both causes. For cause 1, the shape parameter is set equal to 5, and scale equal to $2.5\exp(0.6X)$, that is, the cause-specific hazard model for cause 1 follows from the Weibull distribution $Wei(5, 2.5\exp(0.6X))$. For cause 2, the cause-specific hazard model follows from $Wei(5.5, 2.2\exp(0.5X))$. This characterises an aging process whereby the rate at which the event occurs increases with time for all individuals. Furthermore, higher values of X can be regarded as an indicator for good health such that those individuals with a high covariate value are more likely to experience the event at a later time. Since any specifications imposed on cause-specific hazard models are testable with competing risks data, the consequences of misspecification of the cause-specific hazards are not considered here.

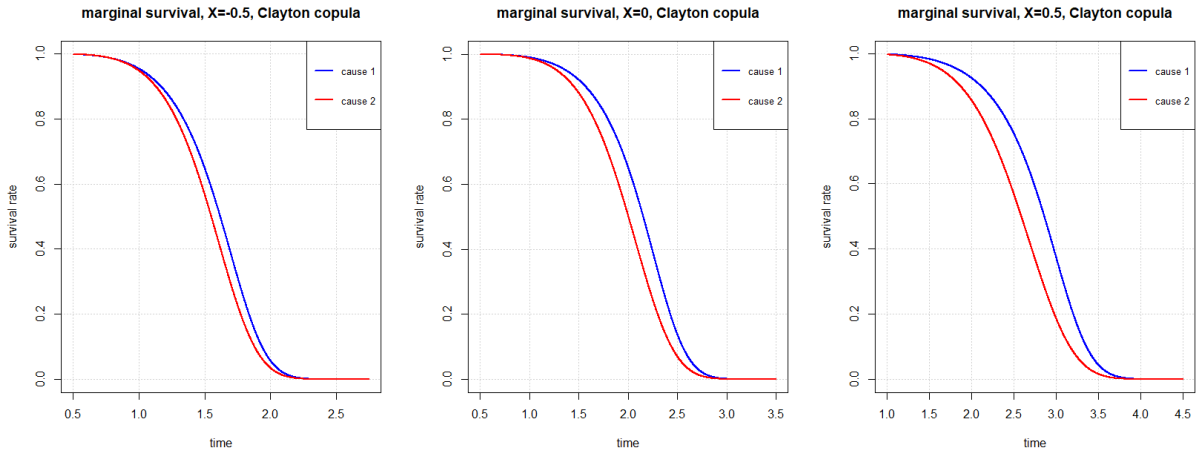


Figure 1. True conditional marginal survival functions with Weibull cause-specific hazards. $T^{(1)}$ and $T^{(2)}$ are related by the dependence structure implied by the Clayton copula with parameter $\alpha = 0.5$.

For modelling dependence between $T^{(1)}$ and $T^{(2)}$, the Frank and Clayton copulas are considered. The copula parameters for the Frank and Clayton copulas are set at $\alpha = 1$ and $\alpha = 0.5$, respectively. This leads to an approximate value of Kendall's tau of 0.670 for the Frank copula and 0.706 for the Clayton copula.

The dependence structures implied by the Frank and Clayton copulas are discussed in Example 2.1. The parameters θ_{j0} of a Weibull cause-specific hazard model $h_j(u, x; \theta_{j0})$ are estimated by regression, $(j = 1, 2)$. Then, the CG estimator of the conditional marginal survival function and corresponding CG confidence intervals are constructed as described in equation (2.5) and Section 3.3.

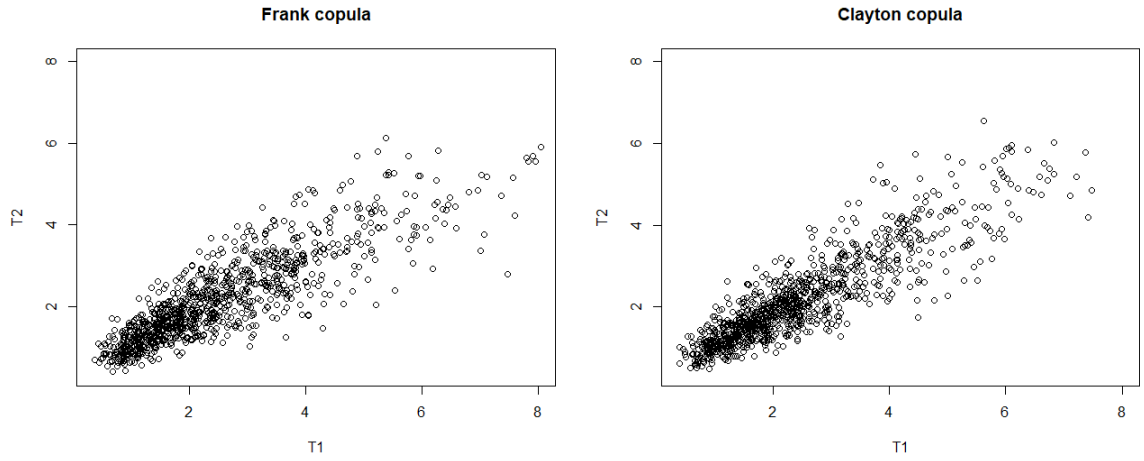


Figure 2. 1000 latent event times generated by the Weibull cause-specific hazard model discussed above and (i) the Frank copula with parameter $\alpha = 1$ (left); (ii) the Clayton copula with parameter $\alpha = 0.5$ (right).

4.2 Bias and coverage under correct specification

The experiment consists of 250 simulations, for each covariate value $X = -0.5$, 0 and 0.5, and for sample sizes $n = 250$, 500, and 750. Let NP denote the (infeasible) nonparametric kernel conditional marginal survival function estimator the full event times $\{T_i^{(1)}, T_i^{(2)}\}_{i=1}^n$, and NP2 denote the nonparametric kernel conditional marginal survival function estimator based on observable competing risks data $\{T_i, \delta_i\}_{i=1}^n$.

The bias results for the three estimators CG, NP and NP2 report the difference between the estimated survival rate and the true survival rate at time t for covariate $X = x$, averaged over the 250 simulations. The coverage results report the proportion that the true survival rate at time t for covariate $X = x$ is contained in the CG confidence interval over the 250 simulations. We also note the proportion the nonparametric survival estimates (NP and NP2) are contained in the CG confidence interval.

In order to learn how tight the CG confidence intervals are, the average width of the interval for covariate $X = x$ and time t is displayed along with coverage results in the same figures.

Only the results for the marginal survival curves for cause 1 are reported; the results for cause 2 were almost identical. Further simulation results that are not presented in this section are given in the appendix.

4.2.1 Frank copula

The bottom row of Figure 3 shows that as the sample size increases, the bias of the CG and NP estimators of the conditional marginal survival function decreases. The nonparametric estimator NP2 that is based on competing risks data is hugely biased; estimated survival rates are biased by over 0.25 at time $t = 1.5$ and $X = -0.5$. The biases of the CG estimator are also significantly lower than the bias from the infeasible NP estimator based on the full sample, especially in small samples.

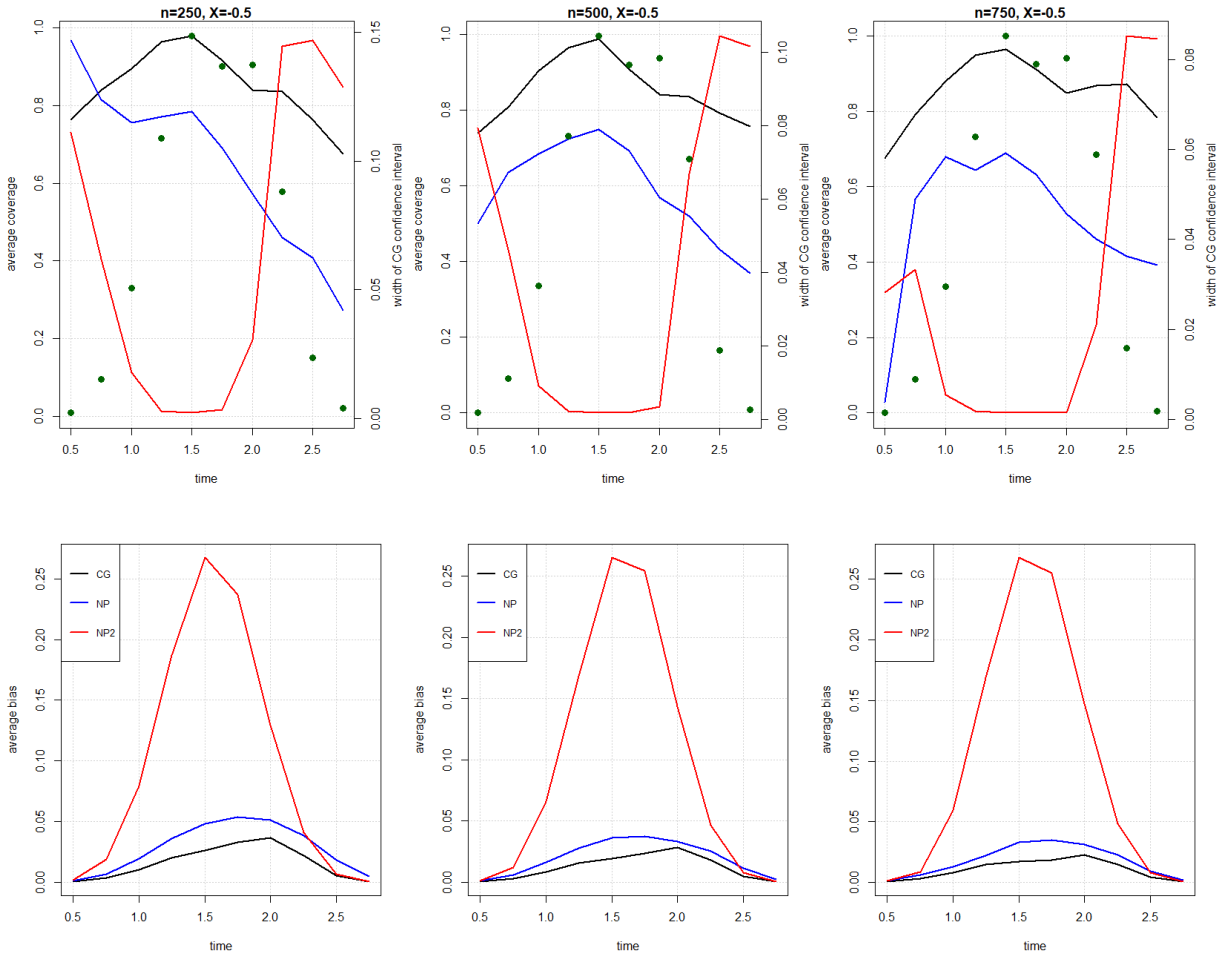


Figure 3. Bias and coverage results for the Frank copula and $X = -0.5$. Black line: CG; blue line: NP; red line: NP2; green dots: width of CG confidence interval. Only the green dots (width of CG confidence intervals) relate to the right-sided axis. The left column of graphs show results for $n = 250$, the middle column for $n = 500$, and the right column for $n = 750$.

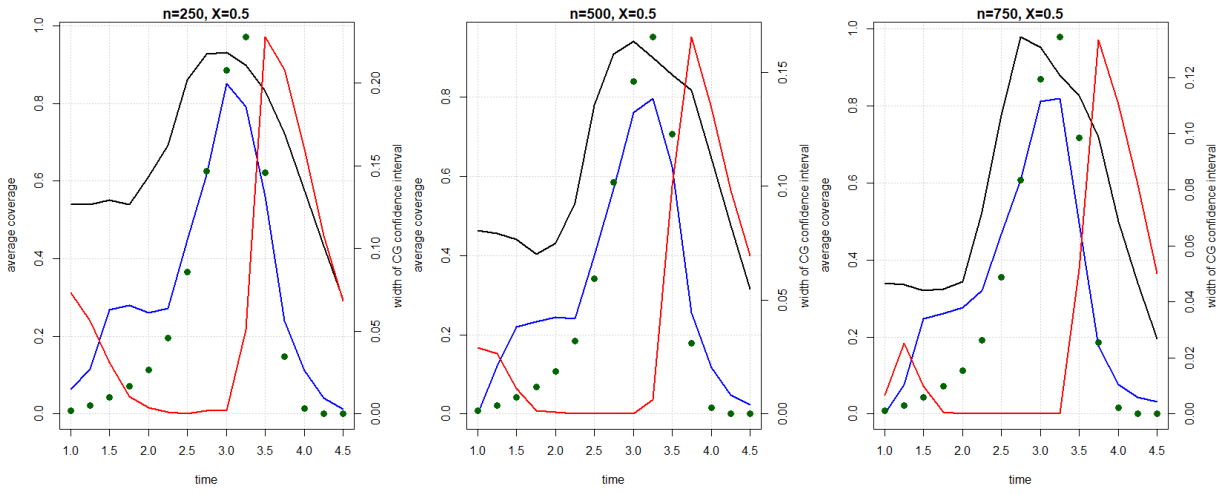
The black lines from the top row of Figure 3 show that coverage of the true survival rates remain relatively stable across the sample size, with a slight fall in coverage as the sample size increases from $n = 500$ to $n = 750$. However, it is important to note the confidence intervals become considerably tighter as the sample size increases, as shown by the green dots. Overall, for the more common survival times between $1 \leq t \leq 2$, coverage is around 90%, 5% down from the nominated coverage of 95%. Coverage is also relatively promising for the rare events which concern very low and very high survival rates.

The rate at which the NP estimate of survival rates is contained within the CG confidence interval falls as the sample size increases (in the top row of Figure 3, the blue curve is significantly below the black curve for large enough sample sizes). Given the high coverage rates of true survival and NP's competitive bias, this suggests the NP estimator has a high variance. The results for $X = 0$ and $X = 0.5$ are reported in the appendix.

4.2.2 Clayton copula

In contrast with the results for the Frank copula, under the Clayton copula the bias of CG is not uniformly lower than NP, although the bottom row of Figure 4 shows that the biases of both CG and NP decrease as the sample size increases.

The top row of Figure 4 presents the coverage results. The right-sided graph in Figure 1 suggests that for cause 1 the survival rate at $t = 2.5$ is around 75% but by $t = 3.5$ the survival rate drops to around 7%. Thus, coverage is competitive for common survival times; for $n = 750$, the true survival rates between $t = 2.5$ and $t = 3.5$ are contained in the CG confidence intervals in over 80% of the experiments.



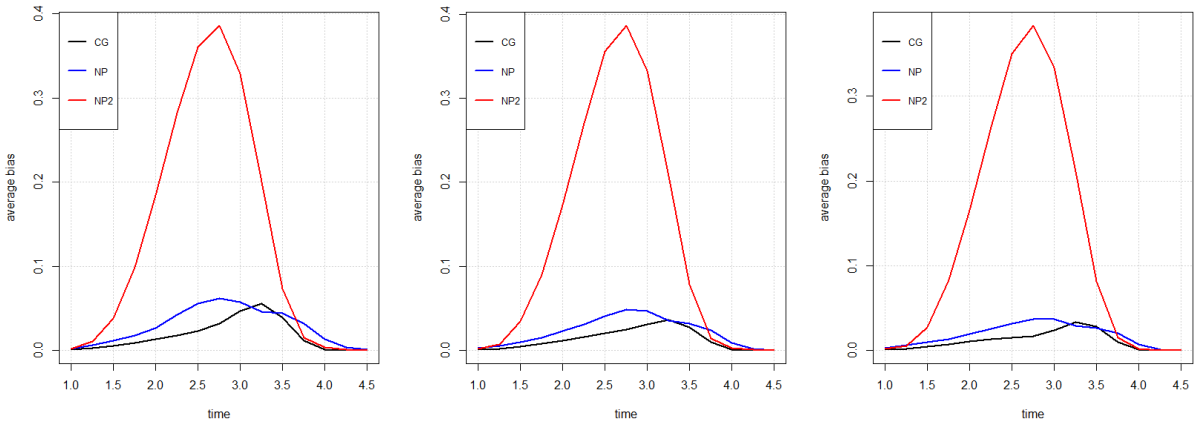


Figure 4. Bias and coverage results for the Clayton copula and $X = 0.5$. Black line: CG; blue line: NP; red line: NP2; green dots: width of CG confidence interval. Only the green dots (width of CG confidence intervals) relate to the right-sided axis. The left column of graphs show results for $n = 250$, the middle column for $n = 500$, and the right column for $n = 750$.

However, coverage of the tails of the true marginal survival function degrade significantly. Invariably, poor coverage of the true survival rate (where the black line is low) is linked with very low widths of the CG confidence interval.

For both the Frank and Clayton copulas, the disastrous results of NP2 warn against ignoring the dependence structure of competing risks. The results confirm that unless the competing risks are independent, the estimated marginal survival rates will be massively biased.

4.3 Misspecification of the copula parameter

Since the copula parameter α is assumed to be known, it is important to consider the impact that misspecification of the parameter value may have for inference. The analysis of the performance of CG confidence intervals in Section 4.2 is repeated for cases in which a wrong copula parameter value is assumed. In particular, the coverage results below show the rate at which the true survival rates for $X = 0$ are contained within the CG confidence intervals.

4.3.1 Frank copula

The top row of Figure 5 shows the sensitivity of bias results to the choice of Frank copula parameter value. The black lines represent the case of correct specification. For small sample sizes, bias appears to be less influenced by the choice of copula parameter, but for $n = 750$ consequences of misspecification are more apparent, especially for the case where $\alpha = 2$.

Similarly to the bias results, coverage is not uniformly higher when the choice of copula parameter is correct. Larger values of α appear to boost coverage when considering survival

rates at lower elapsed times $t \leq 2$, whereas lower values of α perform better for higher elapsed times.

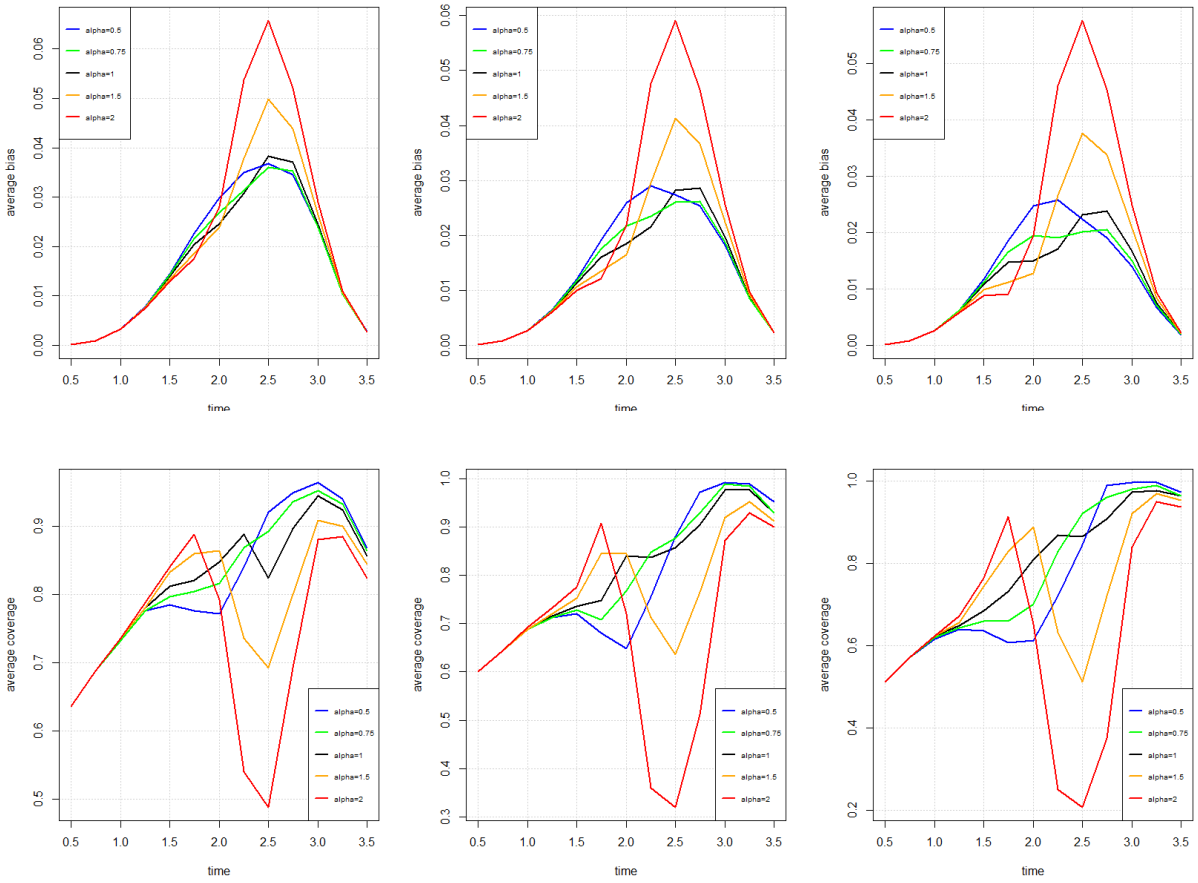


Figure 5. Sensitivity of bias and coverage results under the Frank copula to different values of the copula parameter. The copula parameter under correct specification is $\alpha = 1$. The left column of graphs show results for $n = 250$, the middle column for $n = 500$, and the right column for $n = 750$.

4.3.2 Clayton copula

The top row of Figure 6 suggests that for the Clayton copula, choices of the copula parameter α that are lower than the true value $\alpha = 0.5$ are less harmful in terms of bias than choosing higher values of α; in fact, $\alpha = 0.35$ appears to perform best, although no choice of α leads to uniformly lower bias.

Similar to the observed pattern for the Frank copula, higher values of the copula parameter lead to higher coverage rates when considering survival at lower elapsed times $t \leq 2$, and worse coverage at $t \geq 2$. This trade-off shows the consequences of assuming a stronger or weaker correlation structure between competing risks; it would be interesting to observe whether this pattern holds more generally for CG estimation with small sample sizes across other copula families.

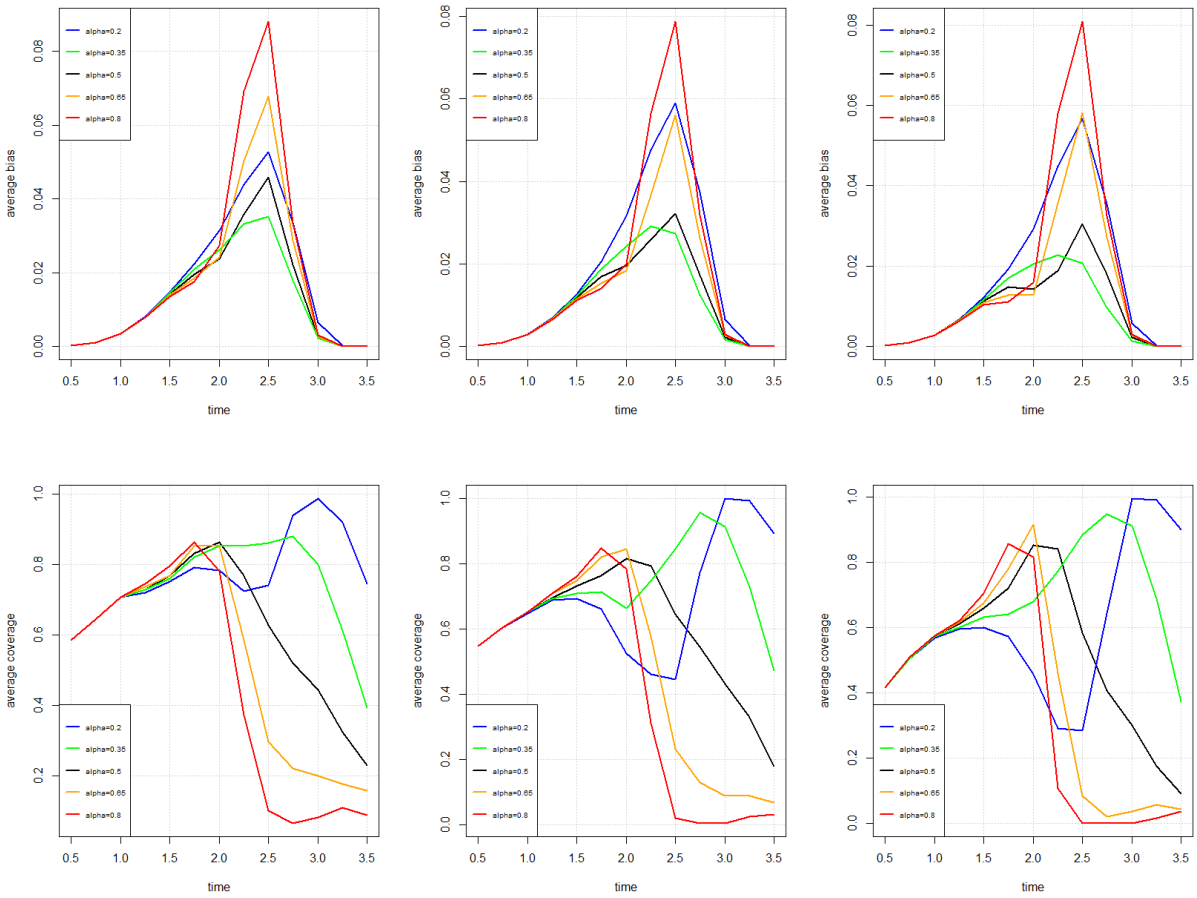


Figure 6. Sensitivity of bias and coverage results under the Clayton copula to different values of the copula parameter. The copula parameter under correct specification is $\alpha = 0.5$. The left column of graphs show results for $n = 250$, the middle column for $n = 500$, and the right column for $n = 750$.

The asymptotic confidence intervals derived here are based on first-order Taylor approximations (see Proof of Corollary 3.1 in the appendix). By equation (2.5), it is clear that the conditional marginal survival is expected to be a highly non-linear function of θ_{j0} (the parameters relating to cause-specific hazard models). In such situations, employing asymptotic approximations for small sample inference may be a concern and therefore, to a certain extent, biases and low coverage rates are to be expected.

From this simulation study it can be seen that the low coverage rates are linked with very low widths of the confidence intervals around the boundary. In order to obtain good coverage when survival rates are close to 0 or 1, the confidence intervals may need to be adapted to make them suitably cautious for use in finite samples.

The results also show that coverage rates are sensitive to the choice of copula parameter; for both the Frank and Clayton copulas, coverage rates appear to be a decreasing function of the copula parameter value as we consider higher elapsed survival times. A larger scale simulation

study would provide more insight into the potential trade-offs between assuming stronger or weaker correlation structures for CG estimators.

5 Conclusion

This paper derives the asymptotic distribution of CG estimators of the conditional marginal survival function for a class of parametric cause-specific hazard models. Using these results, we consider the use of CG-estimated asymptotic confidence intervals for finite sample inference on conditional marginal survival functions with competing risks data. The simulation study suggests the finite sample performance of the confidence intervals may depend heavily on the dependence structure and the area of the survival function that is being considered. The choice of copula parameter also appears to be important for good coverage.

There are many extensions that merit further investigation. It is important to investigate the performance of CG-estimated confidence intervals for a wider class of copula functions within the Archimedean family. For example, more complex copula specifications, such as the Joe-Clayton copula, may be helpful for modelling empirically-relevant problems.

Furthermore, in many settings, researchers may be aware of certain characteristics of the relationship between competing risks, however specifying a copula family may be undesirable. A more agnostic approach may avoid the specification of functional forms for the copula between risks, and instead derive bounds on the conditional marginal survival function based on partial identification methods. For example, Arellano and Bonhomme ([2], pp.6-7) consider partial identification bounds for quantile analysis based on worst-case Frèchet bounds on the copula.

References

[1] Cande V. Ananth and David G. Kleinbaum. Regression models for ordinal responses: A review of methods and applications. *International Journal of Epidemiology*, 26(6):1323–1333, 1997.

[2] Manuel Arellano and Stéphane Bonhomme. Quantile Selection Models With an Application to Understanding Changes in Wage Inequality. *Econometrica*, 85(1):1–28, 2017.

[3] Roel Braekers and Noel Veraverbeke. A copula-graphic estimator for the conditional survival function under dependent censoring. *The Canadian Journal of Statistic*, 33(3):429–447, 2005.

[4] Adrian Colin Cameron and Pravin K. Trivedi. *Microeconometrics - Methods and Applications*. 2005.

[5] Lajmi Lakhal Chaieb, L.-P. Rivest, and Belkacem Abdous. Estimating survival under a dependent truncation. *Biometrika*, 93(3):655–669, sep 2006.

[6] Yi Hau Chen. Semiparametric marginal regression analysis for dependent competing risks under an assumed copula. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 72(2):235–251, 2010.

[7] David R. Cox. *Renewal Theory*. Methuen, 1962.

[8] Paul Embrechts. Copulas: A personal view. *Journal of Risk and Insurance*, 76(3):639–650, 2009.

[9] Jason P Fine and Robert J Gray. A Proportional Hazards Model for the Subdistribution of a Competing Risk Stable. *Journal of the American Statistical Association*, 94(446):496–509, 1999.

[10] Thomas. R Fleming and David P Harrington. *Counting Processes and Survival Analysis*. Wiley, 1991.

[11] Tristen Hayfield and Jeffrey S. Racine. Nonparametric Econometrics: The np Package. *Journal of Statistical Software*, 27(5), 2008.

[12] Jong Hyeon Jeong and Jason Fine. Direct parametric inference for the cumulative incidence function. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 55(2):187–200, 2006.

[13] Jong Hyeon Jeong and Jason P. Fine. Parametric regression on cumulative incidence function. *Biostatistics*, 8(2):184–196, 2007.

[14] John D Kalbfleisch and Ross L Prentice. *Counting Processes and Asymptotic Theory*. Wiley, 2002.

[15] Yi Liu and Qihua Wang. Copula-graphic estimators for the marginal survival function with censoring indicators missing at random. *Statistics and Probability Letters*, 107:101–110, 2015.

[16] Simon M S Lo, Gesine Stephan, and Ralf A Wilke. Competing Risks Copula Models for Unemployment Duration: An Application to a German Hartz Reform. *Journal of Econometric Methods*, 6(1):1–20, 2015.

[17] S Lo, Simon , M and Ralf Wilke, A. A Regression Model for the Copula-Graphic Estimator. *Journal of Econometric Methods*, 3(1):21–46, 2014.

[18] Alexander J. McNeil and Johanna Neslehova. Multivariate Archimedean copulas, d-monotone functions and L1-norm symmetric distributions. *The Annals of Statistics*, 37(5B):3059–3097, 2009.

[19] Whitney K. Newey and Daniel McFadden. Large sample estimation and hypothesis testing. *Handbook of Econometrics*, 4:2111–2245, 1994.

[20] Margaret Sullivan Pepe. Inference for events with dependent risks in multiple endpoint studies. *Journal of the American Statistical Association*, 86(415):770–778, 1991.

[21] Louis-Paul Rivest and Martin T. Wells. A Martingale Approach to the Copula-Graphic Estimator for the Survival Function under Dependent Censoring. *Journal of Multivariate Analysis*, 79(1):138–155, 2001.

[22] A. A. Tsiatis. A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences of the United States of America*, 72(1):20–22, 1975.

[23] Mei-Jie Zhang, Xu Zhang, and Thomas H Scheike. Modelling cumulative incidence function for competing risks data. *Expert Rev Clin Pharmacol.*, 1(3):391–400, 2008.

[24] Ming Zheng and John P. Klein. Estimates of marginal survival for dependent competing risks based on an assumed copula. *Biometrika*, 82(1):127–138, 1995.

A Appendix

A.1 Preliminaries

Some common definitions and results from the literature on counting processes are stated here, see for example, Kalbfleisch and Prentice ([14], Chapter 5) and Fleming and Harrington [10] for general results on counting processes, and Pepe [20] for results applied to the competing risks setting.

Recall from Section 3.1 that $N_{ji}(t)$ is the right-continuous counting process that counts the number of events due to cause j for individual i, and $Y_i(t)$ is the left-continuous at-risk process for individual i. Define the *filtration* (or history)

$$\mathscr{F}_t = \sigma\{N_{ji}(u),\, Y_i(u^+),\, X_i,\, j = 1, 2,\, i = 1, ..., n,\, 0 \leq u \leq t\},\, t \geq 0,$$

where $Y_i(u^+) = \lim_{s \to u^+} Y_i(s)$, and $\sigma[.]$ specifies the sigma algebra of events generated by variables N_{ji}, Y_i and X_i.

Under our model assumptions,

$$\mathbb{P}(dN_{ji}(t) = 1 | \mathscr{F}_t) = Y_i(t) h_j(t, x_i; \theta_{j0}) dt,\ j = 1, 2,\ i = 1, ..., n$$

for $t > 0$ and

$$M_{ji}(t) = N_{ji}(t) - \int_0^t Y_i(u) h_j(u, x_i; \theta_{j0}) du,\ j = 1, 2,\ i = 1, ..., n \tag{A.1}$$

are independent martingales with respect to $\mathscr{F}_t$, see Kalbfleisch and Prentice ([14], Chapter 8, Section 8.2.7, p.265).

The term $\int_0^t Y_i(u) h_j(u, x_i; \theta_{j0}) du$ is called the *compensator* of the counting process N_{ji} with respect to the filtration $\mathscr{F}_t$.

A stochastic process $U(t)$ is *adapted* to $\mathscr{F}_t$ if for each $t \geq 0$, the value of $U(t)$ is a function of $\mathscr{F}_t$.

A stochastic process $U(t)$ is *predictable* with respect to $\mathscr{F}_t$ if for each $t \geq 0$, the value of $U(t)$ is a function of $\mathscr{F}_{t^-}$.

A mean-zero martingale $M(t)$ is said to be *square integrable* if it has a finite variance, that is, if $\mathbb{E}[M^2(t)] < \infty$ for all $t \leq \tau$.

A function f with domain $[0, \infty)$ is *locally bounded* if it is bounded on each interval $[0, s]$, $s < \infty$.

The *predictable variation* process of a square-integrable martingale M is

$$\langle M \rangle(t) = \int_0^t var[dM(u)|\mathscr{F}_{u^-}].$$

For cause j and each i, the predictable variation process is

$$\langle M_{ji} \rangle(t) = \int_0^t Y_i(u) h_j(u, x_i; \theta_{j0}) du,$$

see Kalbfleisch and Prentice ([14], Chapter 5, Section 5.3.2, equation (5.23)).

If $M_{ji}(t)$ is a martingale with respect to $\mathscr{F}_t$ and $H_{ji}(t)$ is a predictable process with respect to $\mathscr{F}_t$, then the process $\{U_{ji}(t),\, 0 \leq t \leq \tau\}$ where $U_{ji}(t) = \int_0^t H_{ji}(u) dM_{ji}(u)$ is a martingale with predictable variation process

$$\langle U_{ji} \rangle(t) = \int_0^t H_{ji}^2(u) dN_{ji}(u),$$

see Kalbfleisch and Prentice ([14], Chapter 5, Section 5.3.2, equation (5.25)).

For $j = 1, 2$, $t \in [0, \tau]$, let

$$U_n = (U_{n1}, U_{n2})' \tag{A.2}$$

be a vector of martingales where

$$U_{jn}(t) = \sum_{i=1}^{n} \int_0^t H_{ji}(u) dM_{ji}(u), \tag{A.3}$$

and for $\epsilon > 0$, define

$$U_{\epsilon jn}(t) = \sum_{i=1}^{n} \int_0^t H_{ji}(u) \mathbb{I}\{|H_{ji}(u)| > \epsilon\} dM_{ji}(u), \tag{A.4}$$

where only jumps of size at least ϵ are taken into account. Since the integrand is a predictable process, $U_{\epsilon nj}(t)$ is a martingale with predictable variation $\langle U_{\epsilon jn} \rangle(t)$.

The conditions for a central limit theorem (CLT) to apply to U_n are now given, see Kalbfleisch and Prentice ([14], Chapter 5, Section 5.5, pp.165-167).

Lemma A.1 (Rebolledo's Theorem - Martingale CLT). *For $j = 1, 2$, suppose $M_{ji}(t)$ for $i = 1, ..., n$ are independent martingales with respect to $\mathscr{F}_t$, and let $H_{ji}(t)$, $i = 1, ..., n$ be predictable functions. For U_n, $U_{jn}(t)$ and $U_{\epsilon jn}(t)$ defined in $(A.2) - (A.4)$, if (i) for any $t \leq \tau$, $\langle U_n \rangle(t) \xrightarrow{p} V(t)$ for some non-random function $V(t)$, and (ii) for all $\epsilon > 0$, $\langle U_{\epsilon jn} \rangle(t) \xrightarrow{p} 0$, then $U_n(t) \xrightarrow{d} \mathcal{N}(0, V(t))$.*

Rebolledo's Theorem gives two conditions that suffice for a CLT to apply to $U_n(t)$. The first requirement is that the covariance of $U_n(t)$ approaches a limit. The second condition ensures

the influence of any single process is negligible in the limit (see Kalbfleisch and Prentice ([14], Chapter 5, p.166).

The next lemma, a variation of Lenglart's inequality, as stated in Fleming and Harrington ([10], Chapter 8, Lemma 8.2.1, p.291), is useful for establishing asymptotic results.

Lemma A.2 (Lenglart's inequality). *Let N be a univariate counting process with continuous compensator A, let $M = N - A$, and let H be a locally bounded, predictable process. Then, for all δ, $\rho > 0$ and any $t \geq 0$,*

$$\mathbb{P}\Big(\sup_{t\in[0,\tau]} \Big| \int_0^t H(u)dM(u) \Big| \geq \rho \Big) \leq \frac{\delta}{\rho^2} + \mathbb{P}\Big(\int_0^\tau H^2(u)dA(u) \geq \delta \Big).$$

Proof. See the proof of Lemma 8.2.1 of Fleming and Harrington ([10], Chapter 8, pp.291-2). □

The next lemma is result on probability bounds.

Lemma A.3. Consider random quantities $E_1, ..., E_S$. For all $S \geq 2$ and $\epsilon > 0$,

$$\mathbb{P}(\sum_{s=1}^S E_s > \epsilon) \leq \sum_{s=1}^S \mathbb{P}\Big(E_s > \frac{\epsilon}{S} \Big).$$

Proof. Note that if the event $\big\{ \sum_{s=1}^S E_s > \epsilon \big\}$ holds then for at least one $s \in \{1, ..., S\}$ the event $\big\{ E_s > \frac{\epsilon}{S} \big\}$ must hold. Therefore, by the union bound,

$$\begin{aligned} \mathbb{P}\Big(\sum_{s=1}^S E_s > \epsilon \Big) &\leq \mathbb{P}\Big(\bigcup_{s=1}^S \big\{ E_s > \frac{\epsilon}{S} \big\} \Big) \\ &\leq \sum_{s=1}^S \mathbb{P}\Big(E_s > \frac{\epsilon}{S} \Big). \end{aligned}$$

□

A.2 Further Notation and Calculations

Some notation to simplify expressions are listed here, along a collection of terms that will be used in derivations.

For $j = 1, 2$, $t \in [0, \tau]$, let $h_{ji}(t; \theta_j) = h_j(t, x_i; \theta_j)$.

The following notation is maintained for the copula generating function: $\phi_\alpha(u) = \phi(u; \alpha)$; $\phi'_\alpha(u) = \partial\phi_\alpha(u)/\partial u$; $\phi''_\alpha(u) = \partial^2\phi_\alpha(u)/\partial u^2$; $\phi_\alpha^{-1(\prime)}(u) = \partial\phi_\alpha^{-1}(u)/\partial u$.

From (2.5),

$$S_j(t, x; \theta, \alpha) = \phi_\alpha^{-1}(A_j(t, x; \theta, \alpha))$$

where $A_j(t,x;\theta,\alpha) := -\int_0^t \phi'_\alpha(S(u,x;\theta))Q'_j(u,x;\theta)du$, and $Q'_j(u,x;\theta) = S(u,x;\theta)h_j(u,x;\theta)$.

For the proof of Corollary 3.1, the following calculations are needed.

Note that

$$\begin{aligned}
\nabla_{\theta_k} S_j(t,x;\theta,\alpha) &= \phi_\alpha^{-1(\prime)}(A_j(t,x;\theta,\alpha))\nabla_{\theta_k} A_j(t,x;\theta,\alpha) \\
\nabla_{\theta_k} A_j(t,x;\theta,\alpha) &= -\int_0^t \phi''_\alpha(S(u,x;\theta))[\nabla_{\theta_k} S(u,x;\theta)]Q'_j(u,x;\theta)du - \int_0^t \phi'_\alpha(S(u,x;\theta)) \\
&\quad \times \nabla_{\theta_k} Q'_j(u,x;\theta)du \\
\nabla_{\theta_k} S(u,x;\theta) &= -\Big(\int_0^u \nabla_{\theta_j} h_k(s,x;\theta_k)ds\Big)S(u,x;\theta)
\end{aligned}$$

and,

$$\begin{aligned}
\nabla_{\theta_j} Q'_j(u,x;\theta) &= -\Big(\int_0^u h_j(s,x;\theta_j)ds\Big)Q'_j(u,x;\theta) + S(u,x;\theta)\nabla_{\theta_j} h_j(u,x;\theta_j). \\
\nabla_{\theta_{k\neq j}} Q'_j(u,x;\theta) &= -\Big(\int_0^u h_k(s,x;\theta_k)ds\Big)Q'_j(u,x;\theta).
\end{aligned}$$

Then,

$$\begin{aligned}
\nabla_{\theta_j} S_j(t,x;\theta,\alpha) &= \phi_\alpha^{-1(\prime)}(A_j(t,x;\theta,\alpha))\Big\{\int_0^t \phi''_\alpha(S(u,x;\theta))[\Big(\int_0^u h_j(s,x;\theta_j)ds\Big)S(u,x;\theta)] \\
&\quad \times Q'_j(u,x;\theta)du + \int_0^t \phi'_\alpha(S(u,x;\theta))\Big[\Big(\int_0^u h_j(s,x;\theta_j)ds\Big)Q'_j(u,x;\theta) \\
&\quad -S(u,x;\theta)\nabla_{\theta_j} h_j(u,x;\theta_j)\Big]du\Big\}
\end{aligned}$$

and

$$\begin{aligned}
\nabla_{\theta_{k\neq j}} S_j(t,x;\theta,\alpha) &= \phi_\alpha^{-1(\prime)}(A_j(t,x;\theta,\alpha))\Big\{\int_0^t \phi''_\alpha(S(u,x;\theta))\Big(\int_0^u h_{k\neq j}(s,x;\theta_{k\neq j})ds\Big)S(u,x;\theta) \\
&\quad \times Q'_j(u,x;\theta)du + \int_0^t \phi'_\alpha(S(u,x;\theta))\Big(\int_0^u h_{k\neq j}(s,x;\theta_k)ds\Big)Q'_j(u,x;\theta)du\Big\}.
\end{aligned}$$

Proof of Lemma 3.1

The proof closely follows Chapter 5 of Kalbfleisch and Prentice [14] and is given here for completeness.

Part (i): Consistency.

From Section 3.1, the full likelihood function (3.1) is

$$L_n(\theta) = \prod_{j=1}^{2}\prod_{i=1}^{n} h_{ji}(t_i;\theta_j)^{\epsilon_{ji}} W_j(t_i,x_i;\theta_j).$$

The log-likelihood is then given by

$$\ell_n(\theta,\tau)=\sum_{j=1}^{2}\sum_{i=1}^{n}\Big(\epsilon_{ji}\log h_{ji}(t_i;\theta_j)-\int_0^{t_i}h_{ji}(u;\theta_j)\Big)du,$$

and this can be written in terms of the counting processes,

$$\ell_n(\theta,\tau) \;=\; \sum_{j=1}^{2}\sum_{i=1}^{n}\int_0^{\tau}\log h_{ji}(u;\theta_j)dN_{ji}(u)-\sum_{j=1}^{2}\sum_{i=1}^{n}\int_0^{\tau}Y_i(u)h_{ji}(u;\theta_j)du.$$

For any $t\in[0,\tau]$, let $v_n(\theta,t)=n^{-1}[\ell_n(\theta,t)-\ell_n(\theta_0,t)]$. That is,

$$v_n(\theta,t) \;=\; \frac{1}{n}\sum_{j=1}^{2}\sum_{i=1}^{n}\int_0^{t}\log\Big(\frac{h_{ji}(u;\theta_j)}{h_{ji}(u;\theta_{j0})}\Big)dN_{ji}(u)-\frac{1}{n}\sum_{j=1}^{2}\sum_{i=1}^{n}\int_0^{t}Y_i(u)[h_{ji}(u;\theta_j)-h_{ji}(u;\theta_{j0})]du.$$

By the model assumption, $\mathbb{P}(dN_{ji}(t)=1|\mathscr{F}_{t^-})=Y_i(t)h_{ji}(t;\theta_{j0})dt$. Let

$$\begin{aligned}A_n(\theta,t) \;=\;& \frac{1}{n}\sum_{j=1}^{2}\sum_{i=1}^{n}\int_0^{t}\log\Big(\frac{h_{ji}(u;\theta_j)}{h_{ji}(u;\theta_{j0})}\Big)Y_i(u)h_{ji}(u;\theta_{j0})du-\frac{1}{n}\sum_{j=1}^{2}\sum_{i=1}^{n}\int_0^{t}Y_i(u)[h_{ji}(u;\theta_j)\\ &-h_{ji}(u;\theta_{j0})]du.\end{aligned}$$

Then,

$$v_n(\theta,t)-A_n(\theta,t)=\frac{1}{n}\sum_{j=1}^{2}\sum_{i=1}^{n}\int_0^{t}\log\Big(\frac{h_{ji}(u;\theta_j)}{h_{ji}(u;\theta_{j0})}\Big)dM_{ji}(u). \tag{A.5}$$

The integrand in $(A.5)$ is locally bounded and predictable by Assumption 3.1, therefore by Theorem 2.3.1 of Fleming and Harrington ([10], Chapter 2, p.61), $v_n(\theta,t)-A_n(\theta,t)$ is a local square integrable martingale, with predictable variation process at t equal to

$$\begin{aligned}\langle\sqrt{n}(v_n(\theta,\cdot)-A_n(\theta,\cdot)),\sqrt{n}(v_n(\theta,\cdot)-A_n(\theta,\cdot))\rangle(t) \;=\;& \frac{1}{n}\sum_{j=1}^{2}\sum_{i=1}^{n}\int_0^{t}\Big[\log\Big(\frac{h_{ji}(u;\theta_j)}{h_{ji}(u;\theta_{j0})}\Big)\Big]^2Y_i(u)\\ &\times h_{ji}(u;\theta_{j0})du\\ =\;& \frac{2}{n}\sum_{j=1}^{2}\sum_{i=1}^{n}\int_0^{t}\log\Big(\frac{h_{ji}(u;\theta_j)}{h_{ji}(u;\theta_{j0})}\Big)Y_i(u)\\ &\times h_{ji}(u;\theta_{j0})du.\end{aligned}$$

Under Assumption 3.1, UWL applies so that for any $u\in[0,\tau]$,

$$\sup_{\theta_j\in\Theta_j}\Big|\frac{1}{n}\sum_{i=1}^{n}\log\Big(\frac{h_{ji}(u;\theta_j)}{h_{ji}(u;\theta_{j0})}\Big)Y_i(u)h_{ji}(u;\theta_{j0})-\mathbb{E}\Big[\log\Big(\frac{h_j(u,x;\theta_j)}{h_j(u,x;\theta_{j0})}\Big)Y(u)h_j(u,x;\theta_{j0})\Big]\Big|\xrightarrow{p}0. \tag{A.6}$$

Also, by boundedness conditions implied by Assumption 3.1,

$$\sup_{u\in[0,\tau],\theta_j\in\Theta_j}\Big|\mathbb{E}\Big[\log\Big(\frac{h_j(u,x;\theta_j)}{h_j(u,x;\theta_{j0})}\Big)Y(u)h_j(u,x;\theta_{j0})\Big]\Big|<\infty.$$

Therefore, $\langle\sqrt{n}(v_n(\theta,\cdot)-A_n(\theta,\cdot)),\sqrt{n}(v_n(\theta,\cdot)-A_n(\theta,\cdot))\rangle(t)$ converges to a finite limit for any $t\in[0,\tau]$. Thus,

$$\langle(v_n(\theta,\cdot)-A_n(\theta,\cdot)),(v_n(\theta,\cdot)-A_n(\theta,\cdot))\rangle(\tau)=o_p(1). \tag{A.7}$$

Using Lemma A.3, and Lemma A.2, Lenglart's inequality, with $\rho=2\kappa^{\frac{1}{4}}$,

$$\begin{aligned}
\mathbb{P}\Big\{\sup_{t\in[0,\tau]}|v_n(\theta,t)-A_n(\theta,t)|>2\kappa^{\frac{1}{4}}\Big\} &\le \sum_{j=1}^{2}\mathbb{P}\Big\{\sup_{t\in[0,\tau]}\Big|\frac{1}{n}\sum_{i=1}^{n}\int_0^t\log\Big(\frac{h_{ji}(u;\theta_j)}{h_{ji}(u;\theta_{j0})}\Big)dM_{ji}(u)\Big|>\kappa^{\frac{1}{4}}\Big\}\\
&\le \frac{\kappa^{\frac{1}{2}}}{4}+\mathbb{P}\Big\{\frac{1}{n^2}\sum_{j=1}^{2}\sum_{i=1}^{n}\int_0^\tau\Big[\log\Big(\frac{h_{ji}(u;\theta_j)}{h_{ji}(u;\theta_{j0})}\Big)\Big]^2Y_i(u)\\
&\quad\times h_{ji}(u;\theta_{j0})du>\kappa\Big\}\\
&\le \frac{\kappa^{\frac{1}{2}}}{4}+o(1),
\end{aligned}$$

where the last inequality follows by $(A.7)$. For small enough κ, it follows that

$$\sup_{t\in[0,\tau]}|v_n(\theta,t)-A_n(\theta,t)|=o_p(1). \tag{A.8}$$

Let

$$v(\theta,\tau)=\sum_{j=1}^{2}\int_0^t\Big(\mathbb{E}\Big[\log\Big(\frac{h_j(u,x;\theta_j)}{h_j(u.x;\theta_{j0})}\Big)Y(u)h_j(u,x;\theta_{j0})\Big]-\mathbb{E}\Big[\int_0^tY(u)[h_j(u,x;\theta_j)-h_j(u,x;\theta_{j0})]\Big]\Big)du.$$

By Assumption 3.1 and $(A.6)$,

$$A_n(\theta,\tau)\xrightarrow{p}v(\theta,\tau). \tag{A.9}$$

Thus, by T,

$$v_n(\theta,\tau)\xrightarrow{p}v(\theta,\tau).$$

Let $v(\theta)=v(\theta,\tau)$. By Assumption 3.1(x),

$$\nabla_{\theta_j}v(\theta)=\int_0^\tau\mathbb{E}\Big[\Big(\frac{\nabla_{\theta_j}h_j(u;x,\theta_j)}{h_j(u;x,\theta_j)}\Big)Y(u)h_j(u;x,\theta_{j0})\Big]du-\int_0^\tau\mathbb{E}\big[Y(u)\nabla_{\theta_j}h_j(u;x,\theta_j)\big]du.$$

Therefore, $\nabla_{\theta_j}v(\theta_0)=0$, $(j=1,2)$.

Also,

$$\begin{aligned}
\nabla_{\theta_j\theta_j'} v(\theta) &= \int_0^\tau \mathbb{E}\Big[\frac{h_j(u,x;\theta_j)[\nabla_{\theta_j\theta_j'} h_j(u,x;\theta_j)] - (\nabla_{\theta_j} h_j(u,x;\theta_j))^{\otimes 2}}{h_j(u,x;\theta_j)} Y(u) h_j(u,x;\theta_{j0})\Big] du \\
&\quad - \int_0^\tau \mathbb{E}\big[Y_j(u)\nabla_{\theta_j\theta_j'} h_j(u,x;\theta_j)\big] du \\
&= \int_0^\tau \mathbb{E}\Big[\Big(\frac{\nabla_{\theta_j\theta_j'} h_j(u,x;\theta_j)}{h_j(u,x;\theta_j)}\Big) Y(u)\{h_j(u,x;\theta_{j0}) - h_j(u,x;\theta_j)\}\Big] du \\
&\quad - \int_0^\tau \mathbb{E}\Big[\Big(\frac{\nabla_{\theta_j} h_j(u,x;\theta_j)}{h_j(u,x;\theta_j)}\Big)^{\otimes 2} Y(u) h_j(u,x;\theta_{j0})\Big] du \\
&:= -\Sigma_j(\theta_j,\tau).
\end{aligned}$$

By Assumption 3.1(ix), $\Sigma_j(\theta_j,\tau)$ is positive definite for all $\theta_j \in \Theta_j$. Also note that $\nabla_{\theta_1\theta_2'} v(\theta) = \nabla_{\theta_2\theta_1'} v(\theta) = 0$. Thus, $\Sigma(\theta,\tau) = \mathrm{diag}(\Sigma_1(\theta_1,\tau), \Sigma_2(\theta_2,\tau))$ is positive definite for all $\theta \in \Theta$. From convexity of $v(\theta)$, $\hat{\theta} \xrightarrow{p} \theta_0$ (cf. Kalbfleisch and Prentice [14], bottom of p.175). This shows part (i) of Lemma 3.1.

Part (ii): Asymptotic Normality

The focus is on estimation of θ_j, $j \in \{1,2\}$.

The score process of cause j based on data available up to a specified time $t \in (0,\tau]$ is

$$U_{jn}(\theta_j,t) = \sum_{i=1}^n \int_0^t \Big(\frac{\nabla_{\theta_j} h_{ji}(u;\theta_j)}{h_{ji}(u;\theta_j)}\Big) dN_{ji}(u) - \sum_{i=1}^n \int_0^t Y_i(u)\nabla_{\theta_j} h_{ji}(u;\theta_j) du,$$

The compensator of $N_{ji}(t)$ is $A_{ji}(t) = \int_0^t Y_i(u) h_{ji}(u;\theta_{j0}) du$, so that $M_{ji}(t) = N_{ji}(t) - A_{ji}(t)$ is a mean-zero martingale with respect to the filtration $\mathscr{F}_t$. Note that

$$\int_0^t \Big(\frac{\nabla_{\theta_j} h_{ji}(u;\theta_{j0})}{h_{ji}(u;\theta_{j0})}\Big) dA_{ji}(u) = \int_0^t Y_i(u)\nabla_{\theta_j} h_{ji}(u;\theta_{j0}) du.$$

Therefore,

$$U_{jn}(\theta_{j0},t) = \sum_{i=1}^n \int_0^t \Big(\frac{\nabla_{\theta_j} h_{ji}(u;\theta_{j0})}{h_{ji}(u;\theta_{j0})}\Big) dM_{ji}(u).$$

The i^{th} term in the sum is a stochastic integral of a predictable vector process with respect to a martingale. $M_{j1},...,M_{jn}$ are independent martingales and $n^{-\frac{1}{2}}\nabla_{\theta_j} \log h_{ji}(u;\theta_{j0})$ is a d_{θ_j}-dimensional vector of predictable processes with respect to the filtration $\mathscr{F}_t$.

Thus, the score process $U_{jn}(\theta_{j0},t)$ is a martingale and its predictable variation process is

$$\langle n^{-\frac{1}{2}} U_{jn}(\theta_{j0})\rangle(t) = \int_0^t \frac{1}{n}\sum_{i=1}^n \Big[\frac{(\nabla_{\theta_j} h_{ji}(u;\theta_{j0}))^{\otimes 2}}{h_{ji}(u;\theta_{j0})}\Big] Y_i(u) du.$$

Under Assumption 3.1, WLLN within the integral applies. That is,

$$\langle n^{-\frac{1}{2}} U_{jn}(\theta_{j0})\rangle(t) \xrightarrow{p} \int_0^t \mathbb{E}\Big[\frac{(\nabla_{\theta_j} h_j(u,x;\theta_{j0}))^{\otimes 2}}{h_j(u,x;\theta_{j0})} Y(u)\Big] du.$$

This satisfies condition (i) of Lemma A.1 (Rebolledo's CLT).

Let θ_{jk} denote the k-th element of θ_j, $(k = 1, ..., d_{\theta_j})$. Construct $U_{\epsilon jn}(t)$ from $(A.4)$ with $H_{ji}(u) = n^{-\frac{1}{2}} \nabla_{\theta_{jk}} \log h_{ji}(u;\theta_{j0})$ and $M_{ji}(u) = N_{ji}(u) - \int_0^u Y_i(s) h_{ji}(s;\theta_{j0}) du$. Since the integrand in $U_{\epsilon jn}(t)$ is a predictable process,

$$\begin{aligned}\langle U_{\epsilon jn}\rangle(t) &= \frac{1}{n}\sum_{i=1}^n \int_0^\tau \Big(\nabla_{\theta_{jk}} \log h_{ji}(u;\theta_{j0})\Big)^2 \mathbb{I}\Big\{\Big|n^{-\frac{1}{2}} \nabla_{\theta_{jk}} \log h_{ji}(u;\theta_{j0})\Big| > \epsilon\Big\} Y_i(u) h_{ji}(u;\theta_{j0}) \\ &\xrightarrow{p} 0,\end{aligned}$$

where the second line holds by Assumption 3.2 for any $k = 1, ..., d_{\theta_j}$, $j = 1, 2$ and $\epsilon > 0$. This is sufficient to satisfy condition (ii) of Lemma A.1; also see Kalbfleisch and Prentice ([14], p.180).

Thus, by Lemma A.1,

$$n^{-\frac{1}{2}} U_{jn}(\theta_{j0}, \tau) \xrightarrow{d} \mathcal{N}\big(0, \Sigma_j(\theta_{j0})\big),$$

where $\Sigma_j(\theta_{j0}) = \Sigma_j(\theta_{j0}, \tau)$.

Note that $\nabla_{\theta_j \theta_j'} \ell_n(\theta, \tau) = \nabla_{\theta_j} U_{jn}(\theta_j, \tau)$, where

$$\begin{aligned}\nabla_{\theta_j} U_{jn}(\theta_j, \tau) &= \sum_{i=1}^n \int_0^\tau \Big(\frac{\nabla_{\theta_j \theta_j'} h_{ji}(u;\theta_j)}{h_{ji}(u;\theta_j)} - \Big(\frac{\nabla_{\theta_j} h_{ji}(u;\theta_j)}{h_{ji}(u;\theta_j)}\Big)^{\otimes 2}\Big) dN_{ji}(u) \\ &\quad - \sum_{i=1}^n \int_0^\tau \nabla_{\theta_j \theta_j'} h_{ji}(u;\theta_j) Y_i(u) du.\end{aligned}$$

By the definition of $M_{ji}(u)$, re-write

$$\begin{aligned}\nabla_{\theta_j} U_{jn}(\theta_j, \tau) &= \sum_{i=1}^n \int_0^\tau \frac{\nabla_{\theta_j \theta_j'} h_{ji}(u;\theta_j)}{h_{ji}(u;\theta_j)} dM_{ji}(u) + \sum_{i=1}^n \int_0^\tau \frac{\nabla_{\theta_j \theta_j'} h_{ji}(u;\theta_j)}{h_{ji}(u;\theta_j)} Y_i(u) h_{ji}(u;\theta_{j0}) du \\ &\quad - \sum_{i=1}^n \int_0^\tau \Big(\frac{\nabla_{\theta_j} h_{ji}(u;\theta_j)}{h_{ji}(u;\theta_j)}\Big)^{\otimes 2} dN_{ji}(u) - \sum_{i=1}^n \int_0^\tau \nabla_{\theta_j \theta_j'} h_{ji}(u;\theta_j) Y_i(u) du.\end{aligned}$$

For any $t \in (0, \tau]$, by Assumption 3.1, UWL, for any consistent estimator $\bar{\theta}_j$ for θ_{j0},

$$\frac{1}{n}\sum_{i=1}^n \frac{\nabla_{\theta_j \theta_j'} h_{ji}(u;\bar{\theta}_j)}{h_{ji}(u;\bar{\theta}_j)} Y_i(u) h_{ji}(u;\theta_{j0}) \xrightarrow{p} \mathbb{E}\big[\nabla_{\theta_j \theta_j'} h_j(u,x;\theta_{j0}) Y(u)\big].$$

Similarly,

$$\frac{1}{n}\sum_{i=1}^{n}\nabla_{\theta_j\theta_j'}h_{ji}(u;\theta_j)Y_i(u)\xrightarrow{p}\mathbb{E}\big[\nabla_{\theta_j\theta_j'}h_j(u,x;\theta_{j0})Y(u)\big].$$

Hence,

$$\begin{aligned}\frac{1}{n}\nabla_{\theta_j}U_{jn}(\bar{\theta}_j,\tau) &= \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau\frac{\nabla_{\theta_j\theta_j'}h_{ji}(u;\bar{\theta}_j)}{h_{ji}(u;\bar{\theta}_j)}dM_{ji}(u)-\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau\Big(\frac{\nabla_{\theta_j}h_{ji}(u;\bar{\theta}_j)}{h_{ji}(u;\bar{\theta}_j)}\Big)^{\otimes 2}dN_{ji}(u)\\ &\quad+o_p(1).\end{aligned}$$

The first term on the RHS is the integral of a predictable process with respect to a martingale, and so has mean zero and is $o_p(1)$. For the second term, by similar arguments to those used above,

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau\Big(\frac{\nabla_{\theta_j}h_{ji}(u;\bar{\theta}_j)}{h_{ji}(u;\bar{\theta}_j)}\Big)^{\otimes 2}dN_{ji}(u)=\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau\Big(\frac{\nabla_{\theta_j}h_{ji}(u;\bar{\theta}_j)}{h_{ji}(u;\bar{\theta}_j)}\Big)^{\otimes 2}Y_i(u)h_{ji}(u;\theta_{j0})du+o_p(1).$$

Also, by Assumption 3.1, UWL, for any consistent estimator $\bar{\theta}_j$ for θ_{j0},

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau\Big(\frac{\nabla_{\theta_j}h_{ji}(u;\bar{\theta}_j)}{h_{ji}(u;\bar{\theta}_j)}\Big)^{\otimes 2}Y_i(u)h_{ji}(u;\theta_{j0})du\xrightarrow{p}\Sigma_j(\theta_{j0}).$$

Finally, consider a Taylor expansion around $\hat{\theta}_j=\theta_{j0}$, for some $\bar{\theta}_j$ on the line segment joining $\hat{\theta}_j$ and θ_{j0},

$$\frac{1}{\sqrt{n}}U(\theta_{j0},\tau)+\frac{1}{n}\nabla_{\theta_j}U(\bar{\theta}_j,\tau)\sqrt{n}(\hat{\theta}_j-\theta_{j0})=0.$$

By the above arguments,

$$\begin{aligned}\sqrt{n}(\hat{\theta}_j-\theta_{j0}) &= \Sigma_j(\theta_{j0})^{-1}\frac{1}{\sqrt{n}}U(\theta_{j0},\tau)+o_p(1) \qquad (A.10)\\ &\xrightarrow{d} \mathcal{N}(0,\Sigma_j(\theta_{j0})^{-1}).\end{aligned}$$

This proves part (ii).

Part (iii): Asymptotic Expansion of CIFs

Recall, $Q_j(t,x;\theta) = h_j(t,x;\theta_j)\prod_{k=1}^2 W_k(t,x;\theta_k)$, $(j=1,2)$. Then,

$$\begin{aligned}\nabla_{\theta_j} Q_j(t,x;\theta) &= \int_0^t \nabla_{\theta_j} h_j(u,x;\theta_j)\prod_{k=1}^2 W_k(u,x;\theta_k)du \\ &\quad + \int_0^t \Big[\nabla_{\theta_j} W_j(u,x;\theta_j)\Big] h_j(u,x;\theta_j)(W_j(u,x;\theta_j))^{-1}\prod_{k=1}^2 W_k(u,x;\theta_k)du \\ &= \int_0^t \Big[\nabla_{\theta_j} h_j(u,x;\theta_j)\Big] S(u,x;\theta)du \\ &\quad - \int_0^t \Big[\int_0^u \nabla_{\theta_j} h_j(s,x;\theta_j)ds\Big] h_j(u,x;\theta_j)S(u,x;\theta)du,\end{aligned}$$

and, for $k \neq j$,

$$\nabla_{\theta_k} Q_j(t,x;\theta) = -\int_0^t \Big[\int_0^u \nabla_{\theta_k} h_k(s,x;\theta_k)ds\Big] h_j(u,x;\theta_j)S(u,x;\theta)du.$$

Let $\nabla_\theta Q_j(t,x;\theta) := \big(\nabla'_{\theta_j} Q_j(t,x;\theta), \nabla'_{\theta_k} Q_j(t,x;\theta)\big)$, and recall $\hat\theta = (\hat\theta_1', \hat\theta_2')'$, $\theta_0 = (\theta_{10}', \theta_{20}')'$. For some $\bar\theta$ on the line segment joining $\hat\theta$ and θ_0, by a Taylor expansion,

$$Q_j(t,x;\hat\theta) = Q_j(t,x;\theta_0) + \Big[\nabla_\theta Q_j(t,x;\bar\theta)\Big]'(\hat\theta - \theta_0).$$

Write the $(d_{\theta_1}+d_{\theta_2})\times(d_{\theta_1}+d_{\theta_2})$ matrix, $\Sigma(\theta_0) = \mathrm{diag}(\Sigma_1(\theta_1), \Sigma_2(\theta_2))$. Then, by consistency of $\hat\theta$ for θ_0, and the delta method,

$$\sqrt{n}(Q_j(t,x;\hat\theta) - Q_j(t,x;\theta_0)) \xrightarrow{d} \mathcal{N}\Big(0, \Big[\nabla_\theta Q_j(t,x;\theta_0)\Big]\Sigma(\theta_0)^{-1}\Big[\nabla_\theta Q_j(t,x;\theta_0)\Big]'\Big).$$

□

Proof of Corollary 3.1

Note that $S_j(t,x;\theta,\alpha)$ is a continuous function of θ. Therefore, consistency of $S_j(t,x;\hat\theta,\alpha)$ for $S_j(t,x;\theta_0,\alpha)$ follows by CMT.

The first-order expansion of $\hat\theta_j$ $(j=1,2)$ is given in $(A.10)$. In particular,

$$\sqrt{n}(\hat\theta_j - \theta_{j0}) = \Sigma_j(\theta_{j0})^{-1}\frac{1}{\sqrt{n}}U_{jn}(\theta_{j0}) + o_p(1).$$

Therefore, by a Taylor expansion around $\hat\theta = \theta_0$,

$$S_j(t,x;\hat\theta,\alpha) - S_j(t,x;\theta_0,\alpha) = \mathcal{H}_{S_j}(t,x;\theta_0)'\begin{pmatrix}\hat\theta_1 - \theta_{10}\\ \hat\theta_2 - \theta_{20}\end{pmatrix} + o_p(n^{-\frac{1}{2}}),$$

where

$$\mathcal{H}_{S_j}(t,x;\theta_0) = \left[\nabla_{\theta_j} S_j(t,x;\theta_0,\alpha)', \nabla_{\theta_{k\neq j}} S_j(t,x;\theta_0,\alpha)'\right]' \tag{A.11}$$

and explicit expressions for $\nabla_{\theta_j} S_j(t,x;\theta_0,\alpha)$ and $\nabla_{\theta_{k\neq j}} S_j(t,x;\theta_0,\alpha)$ are provided in Section A.2 of the Appendix.

By Cramér's theorem,

$$\sqrt{n}(S_j(t,x;\hat{\theta},\alpha) - S_j(t,x;\theta_0,\alpha)) \xrightarrow{d} \mathcal{N}(0, \mathcal{V}_{S_j}(t,x))$$

where

$$\mathcal{V}_{S_j}(t,x) = \mathcal{H}_{S_j}(t,x;\theta_0)'\Sigma_{S_j}\mathcal{H}_{S_j}(t,x;\theta_0) \tag{A.12}$$

and

$$\Sigma_{S_j} = \begin{pmatrix} \Sigma_1(\theta_{10})^{-1} & 0 \\ 0 & \Sigma_2(\theta_{20})^{-1} \end{pmatrix}. \tag{A.13}$$

□

B Further simulation results

B.1 Frank copula: $X = 0$

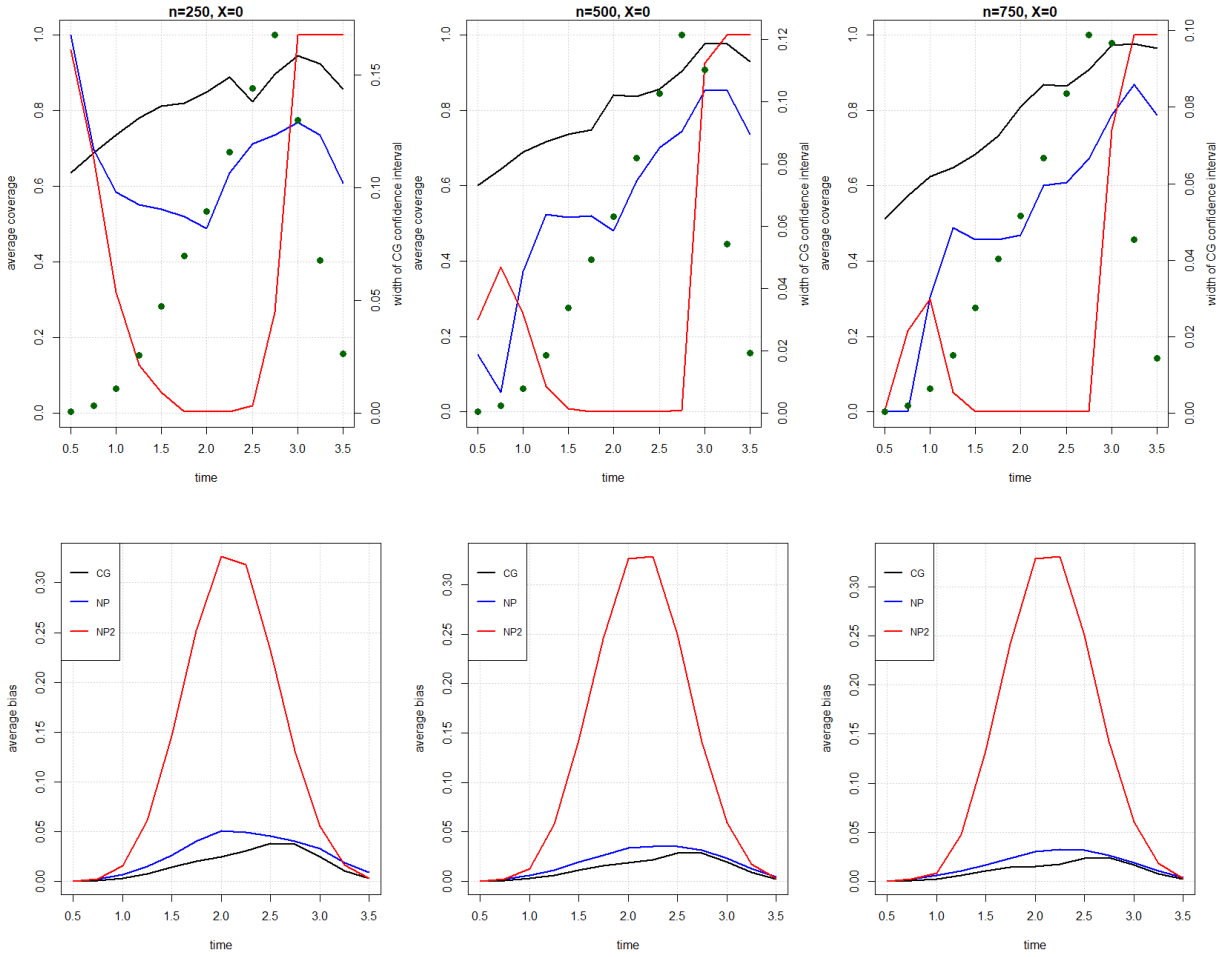


Figure 7. Bias and coverage results for the Frank copula and $X = 0$. Black line: CG; blue line: NP; red line: NP2; green dots: width of CG confidence interval. Only the green dots (width of CG confidence intervals) relate to the right-sided axis. The left column of graphs show results for $n = 250$, the middle column for $n = 500$, and the right column for $n = 750$.

B.2 Frank copula: $X = 0.5$

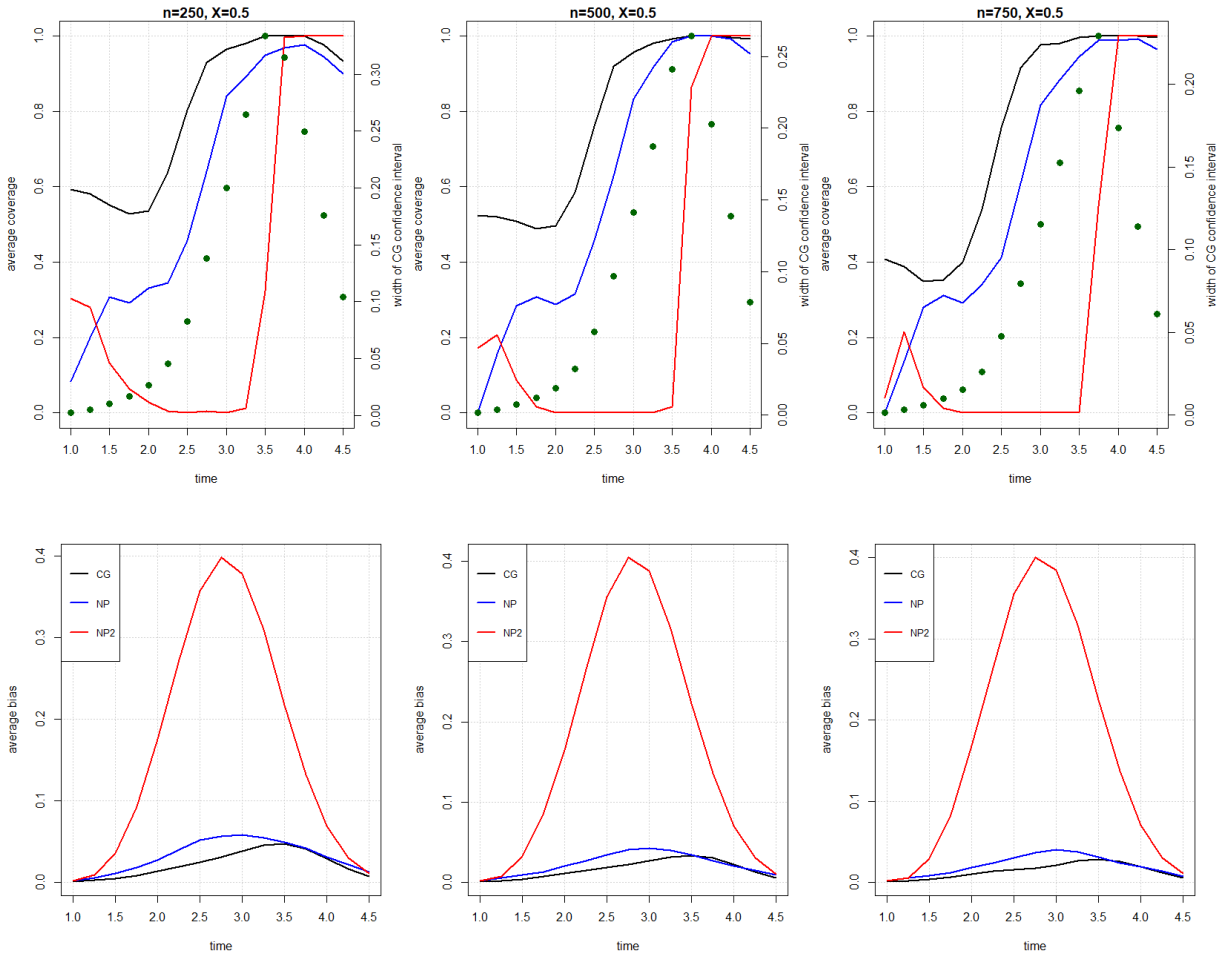


Figure 8. Bias and coverage results for the Frank copula and $X = 0.5$. Black line: CG; blue line: NP; red line: NP2; green dots: width of CG confidence interval. Only the green dots (width of CG confidence intervals) relate to the right-sided axis. The left column of graphs show results for $n = 250$, the middle column for $n = 500$, and the right column for $n = 750$.

B.3 Clayton copula: $X = -0.5$

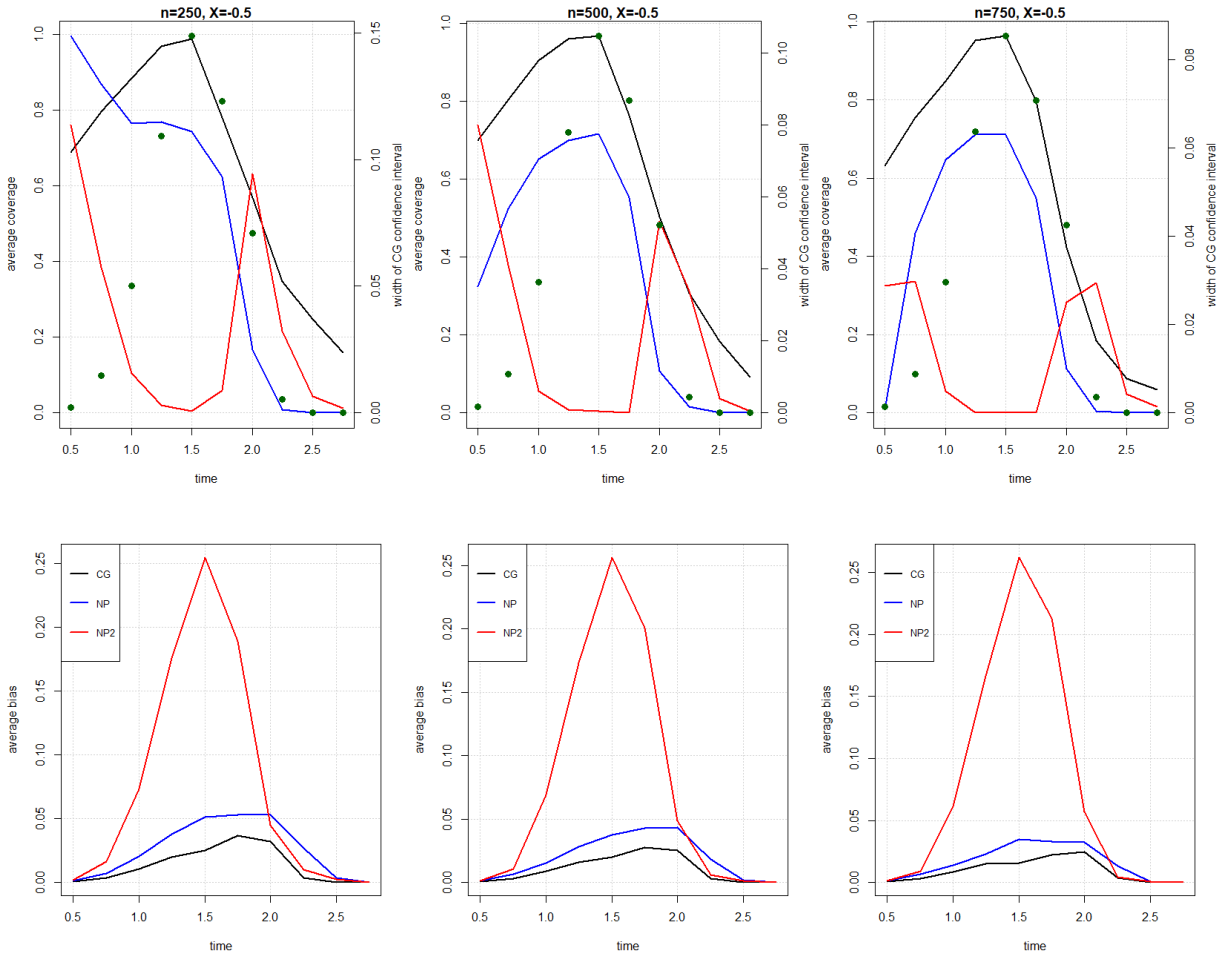


Figure 9. Bias and coverage results for the Clayton copula and $X = -0.5$. Black line: CG; blue line: NP; red line: NP2; green dots: width of CG confidence interval. Only the green dots (width of CG confidence intervals) relate to the right-sided axis. The left column of graphs show results for $n = 250$, the middle column for $n = 500$, and the right column for $n = 750$.

B.4 Clayton copula: $X = 0$

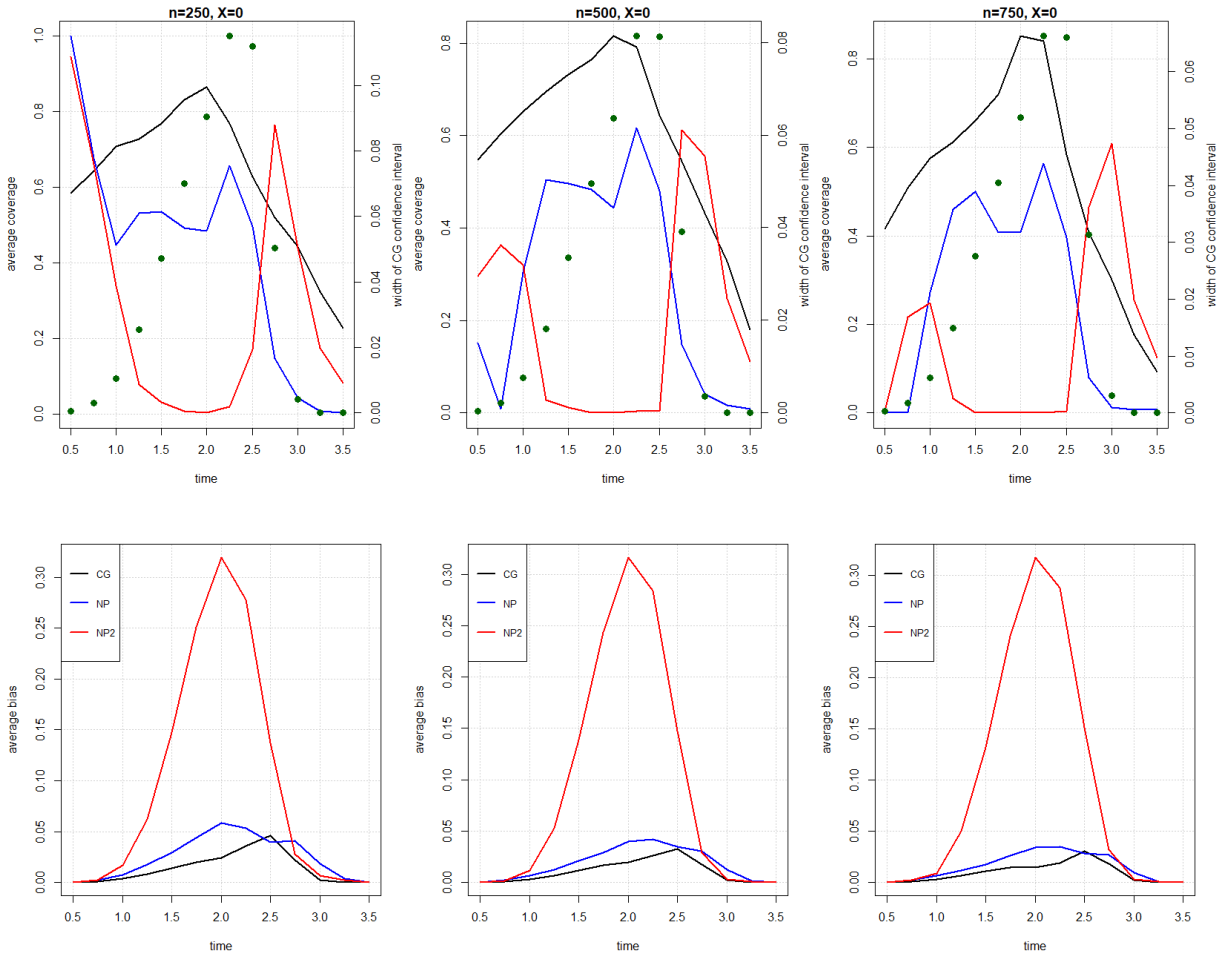


Figure 10. Bias and coverage results for the Clayton copula and $X = 0$. Black line: CG; blue line: NP; red line: NP2; green dots: width of CG confidence interval. Only the green dots (width of CG confidence intervals) relate to the right-sided axis. The left column of graphs show results for $n = 250$, the middle column for $n = 500$, and the right column for $n = 750$.

Ph.D. Thesis: Chapter Four

Large Deviation Bounds for Inference in Moment Condition Models*

Ashish Patel[1] Richard J. Smith[1,2,3,4]

University of Cambridge[1]
cemmap, I.F.S. and U.C.L.[2]
ESCoE[3]
University of Melbourne[4]

Abstract

This paper examines the validity of over-identified moment restrictions when violations may occur only in small subgroups of the population. Hansen's J-test and likelihood-based variants aim to have non-trivial power against a wide range of alternatives, whereas power against particular forms of heterogeneity or parameter instability are often of concern. This paper addresses this issue by providing concentration inequalities designed to detect patterns of model misspecification. For continuously-updating Generalised Method of Moments (Hansen et al. [15]), non-asymptotic bounds are derived for large deviations of estimators and goodness-of-fit statistics. The associated bounds can be used to identify subsets of individual characteristics that are not consistent with the moment restrictions. These results are applied to show the consistency of goodness-of-fit statistics (Ramalho and Smith [30]) with data-dependent partitions.

Keywords: Generalised Empirical Likelihood, Overidentifying Moment Restrictions, Large Deviations, Vapnik-Chervonenkis Inequality

*We thank Martin Anthony, Alexis DeBoeck, Oliver Linton, Alexey Onatskiy and Taisuke Otsu for helpful comments. Ashish Patel gratefully acknowledges that part of this research was undertaken while financially supported by an ESRC Studentship Award. We gratefully acknowledge financial support received by a Keynes Fund research grant (Faculty of Economics, University of Cambridge).

1 Introduction

An important and challenging issue in hypothesis testing concerns the identification of departures from the null hypothesis when a model is violated only by a small subgroup of the population. Detection of this type of heterogeneity is of central importance in cross-sectional models. For example, in demand analysis the parameter of interest may represent preferences that can assist targeting marketing efforts, while for treatment effects models identifying heterogeneous behaviour may inform clinical trial designs.

The true theoretical structures that underpin agent-level heterogeneity may be highly complex; involving nonlinear combinations of several individual characteristics, some of which may be unobservable. Since such hypotheses on such structures are difficult to formulate, omnibus test statistics which have non-trivial power against a wide set of alternatives are often implemented. In the moment conditions setting, Hansen's J-test [14] is such a test statistic.

In particular, the J-test and related variants have asymptotic local power equal to size against local alternatives characterised by parameter variation. Hall ([13], p.5) consequently argues that parameter instability tests should be reported as additional model diagnostics. This paper tackles this issue by providing diagnostic tools for such a purpose.

Tests for overidentifying moment restrictions are evaluated on the basis of asymptotic and finite-sample properties. Kitamura [20] (also see Kitamura et al. [21]) highlights the large-deviation optimality of the empirical likelihood ratio (ELR) test for moment condition models. In particular, in the class of tests for which Type I errors tend to zero, the ELR test maximises the rate at which Type II errors tend to zero. When misspecification is due to neglected heterogeneity, Hahn et al. [12] obtain the linear combination of estimated moment functions required to construct optimal m-tests that maximise the non-centrality parameter of the limiting chi-squared distribution of test statistics in the presence of local parameter heterogeneity. This optimality property, however, does not extend to more involved forms of heterogeneity, that is, non-linear structures.

The finite-sample performance of the partition-based Pearson-type statistic of Ramalho and Smith [30] henceforth, RS, indicates improvements in size properties relative to competing tests over a range of simulation settings. Such tests compare a distribution function estimate which uses information derived from the moment restrictions to the empirical distribution function (EDF). Since the EDF is a consistent estimator of the true distribution, a normalised contrast of the two distribution estimators taken over partitioned sets of the covariate space form a goodness-of-fit test. Further developments of this principle include Otsu and Whang [28] and Otsu et al. [27] for testing conditional moment restrictions, and Guay and Lamarche [10] for testing for the presence of structural breaks.

This paper proposes a novel approach to inference in moment condition models by the development and application of concentration inequalities. The associated test statistics have power

against all alternatives while also being designed to detect complicated patterns of misspecification that may occur. A comparison of generalised empirical likelihood (GEL) and empirical measures (cf. RS tests) over carefully-chosen sets is shown to be particularly useful for detecting departures from the null hypothesis in different population regions. Classification methods can select subsets which highlight potential violations of the null hypothesis; the greater the complexity of the classifier, as measured by the Vapnik-Chervonenkis (VC) dimension, the greater the ability to pick out patterns of deviations away from the null hypothesis.

This paper makes a number of contributions to the literature on inference in moment condition models. An open question for Pearson-type chi-squared testing methods concerns how best to choose partitions of the covariate space that can serve to compare goodness-of-fit over different regions. The paper proposes a selection of subsets according to a goodness-of-fit type loss function. The incorporation of pattern recognition algorithms to improve the ability of test statistics to detect departures from the null hypothesis in the moment conditions setting represents a significant improvement on usual testing procedures. Data-driven classifiers are also shown to optimally identify subsets of the covariate space most inconsistent with the moment condition model which provides new insights as compared to usual portmanteau tests and, thus, are potentially useful from a policy-design perspective.

The results discussed above hold for the general GEL class. The VC bounds studied are non-asymptotic in nature, and for the particular case of continuously-updating GMM (Hansen et al. [15]), explicit constants are derived for the concentration inequalities.

The paper is organised as follows. Section 2 describes the moment conditions framework, GEL estimation and the use of GEL implied probabilities for test construction. Section 3 presents the large deviation bounds and VC inequalities in the moment conditions setting. Section 4 applies the VC bounds for the identification of subsets of the population that are unrepresentative of a moment condition model. Section 5 constructs a goodness-of-fit statistic with data-dependent partitions, with simulation results illustrating its potential efficacy. Section 6 concludes. All proofs are given in the appendix.

The following abbreviations are used. p.d.: positive definite. w.p.a.1: with probability approaching one, $\xrightarrow{p}$: converges in probability to, $\xrightarrow{d}$: converges in distribution to, M: the Markov inequality, T : the triangle inequality, CS: the Cauchy-Schwarz inequality, WLLN: the weak law of large numbers, LIE: the law of iterated expectations, LHS: left hand side, RHS: right hand side. $||.||$ is the Euclidean norm, and $||.||_1$ is the ℓ_1 norm. Throughout the derivations in the Appendix, the property that the Euclidean norm is bounded above by the ℓ_1 norm is used.

2 Inference in moment condition models

2.1 Moment condition model

Let $z \in \mathcal{Z}$ denote a d_z-vector of random variables. Let $g : \mathcal{Z} \times \Theta \to \mathbb{R}^{d_g}$ be a vector of known functions of the data vector z and an unknown, finite-dimensional d_θ-vector of parameters $\theta \in \Theta$, $g(z,\theta) = (g^{(1)}(z,\theta), ..., g^{(d_g)}(z,\theta))'$. For the identification of θ, it is required that $d_g \geq d_\theta$; here only the overidentified case $d_g > d_\theta$ is considered. The moment condition model is

$$\mathbb{E}[g(z,\theta)] = 0 \tag{2.1}$$

where expectation $\mathbb{E}[.]$ is taken with respect to the distribution of z.

The specification (2.1) is sufficiently general to describe a number of well-known economic models. Of particular interest are models that seek to aggregate individual-level behaviour, and thus are at risk of neglecting the possibility that preference parameters may vary across individuals.

Example. 2.1. Nonlinear wage regressions with auxiliary census data (Hellerstein and Imbens [16])
Let y be the recorded wage of an individual and $x_1 \in \mathcal{X}$ be a d_{x_1}-vector of covariate information, such as experience, qualifications and age. Let $x_2 \in \mathcal{X}$ be a d_{x_2}-vector of additional variables (possibly overlapping with x_1) for which, through census records, population moments are known. Then for known functions $h_1 : \mathcal{X} \times \Theta \to \mathbb{R}$ and $h_2 : \mathcal{X} \to \mathbb{R}$, joint estimation of the following moment conditions yields efficient estimators of θ,

$$\mathbb{E}[g(z,\theta)] = \mathbb{E}\begin{bmatrix} y - h_1(x_1,\theta) \\ h_2(x_2) \end{bmatrix} = 0.$$

For Example 2.1 the particular focus of this paper would be on identifying instability in the unknown parameter θ; if a component in the parameter vector θ describes a parameter of interest, such as returns to schooling, the statistics can detect the presence of heterogeneity, where returns to schooling varies across individuals systematically with characteristics (x_1, x_2).

Assumption 2.1. **(i)** Θ *is compact and* $\theta_0 \in int(\Theta)$**; (ii)** *there exists* $C_g > 0$ *such that for all* $k \in \{1, ..., d_g\}$**;** $\sup_{z\in\mathcal{Z}} \sup_{\theta\in\Theta} |g^{(k)}(z,\theta)| \leq C_g$**; (iii)** *there exists* $C_G > 0$ *such that for all* $k \in \{1, ..., d_g\}$, $j \in \{1, ..., d_\theta\}$, $\sup_{z\in\mathcal{Z}} \sup_{\theta\in\Theta} |\partial g^{(k)}(z,\theta)/\partial\theta_j'| \leq C_G$, *and* $G = \mathbb{E}[\partial g(z,\theta_0)/\partial\theta']$ *is of rank* d_θ**; (iv)** $\Omega = \mathbb{E}[g(z,\theta_0)g(z,\theta_0)']$ *is a positive definite (p.d.) matrix with minimum eigenvalue greater than* $\delta_{\Omega,0} > 0$ *, and for all* $\theta \in \Theta$, $\Omega(\theta) = \mathbb{E}[g(z,\theta)g(z,\theta)']$ *is a p.d. matrix with minimum eigenvalue greater than* $\delta_{\Omega,\min} > 0$**; (v)** $g(z,\theta)$ *is continuously differentiable on* Θ *for all* $z \in \mathcal{Z}$**; (vi)** *there exists* $\theta_0 \in \Theta$ *such that* $\mathbb{E}[g(z,\theta_0)] = 0$, *and for all* $\theta \neq \theta_0$, $\mathbb{E}[g(z,\theta)] \neq 0$.

Assumption 2.1 comprises standard conditions required for identification and inference in moment condition models together with some stronger boundedness assumptions required to obtain large deviation bounds. Usually for GMM and GEL, it is assumed that $\mathbb{E}[\sup_{\theta\in\Theta}||g(z,\theta)||^{\alpha}]$ is bounded for $\alpha \geq 2$, which implies $\sup_{\theta\in\Theta}\max_{1\leq i\leq n}||g(z_i,\theta)|| = O_p(n^{\frac{1}{\alpha}})$, see the proof of Theorem 3.1 of Newey and Smith [25], whereas Assumption 2.1(ii) assumes the moment function itself is bounded which is a stronger condition; this could be relaxed to allow the moment function to be bounded up to a rate $O(n^{\alpha})$ for some small $\alpha > 0$.

Assumption 2.1(iv) may be relaxed by assuming $\Omega(\theta)$ is p.d. only on a neighbourhood of θ_0, and requiring various exponential quantities be bounded in expectation, cf. Otsu [26], Assumption 2.5(b). However, such extensions are not considered here. Assumption 2.1(vi) is a standard condition allowing point identification of the parameter θ_0.

Assumption 2.2 concerns various GEL quantities described further in Section 2.2.

Assumption 2.2. **(i)** *the function $\rho(v)$ is strictly concave on $\mathcal{V}$, an open interval that contains 0, $\rho_1(0) = \rho_2(0) = -1$, Λ is compact and $0 \in int(\Lambda)$;* **(ii)** *$\Lambda_n = \{\lambda : ||\lambda|| \leq C_{\lambda,n}\}$ for some scalar function $C_{\lambda,n} > 0$ decreasing in n;* **(iii)** *for each $\theta \in \Theta$, $\lambda(\theta) = \arg\max_{\lambda\in\Lambda}\mathbb{E}[\rho(\lambda' g(z,\theta))]$ belongs to $int(\Lambda_n)$;* **(iv)** *for all $\theta \in \Theta$, there exists $T_1 > 0$ and neighbourhoods $\mathcal{N}_\theta$ and $\mathcal{N}_{\lambda(\theta)}$ of θ and $\lambda(\theta)$ respectively, satisfying $\mathbb{E}[\exp\left(T_1 \sup_{\upsilon\in\mathcal{N}_\theta}\sup_{\lambda\in\mathcal{N}_{\lambda(\theta)}}||\rho_1\ (\lambda' g(z,\upsilon))\frac{\partial g(z,\upsilon)}{\partial\theta'}||\right)] < \infty$;* **(v)** *there exists $T_{10} > 0$ and neighbourhoods $\mathcal{N}_\rho$ and $\mathcal{N}'_\rho$ of θ_0 and 0, respectively, satisfying $\mathbb{E}[\exp\left(T_{10}\sup_{\theta\in\mathcal{N}_\rho}\sup_{\lambda\in\mathcal{N}'_\rho}||\rho_2(\lambda' g(z,\theta))g(z,\theta)g(z,\theta)'||\right)] < \infty$;* **(vi)** *for $\rho_i^{\Omega} = \min_{\theta\in\mathcal{N}_\theta,\lambda\in\mathcal{N}_{\lambda(\theta)}}\rho_2(\lambda' g(z_i,\theta))g(z_i,\theta)g(z_i,\theta)'$, $\frac{1}{n}\sum_{i=1}^n \rho_i^{\Omega}$ has minimum eigenvalue $\delta_\rho > 0$, where $\mathcal{N}_\theta$ and $\mathcal{N}_{\lambda(\theta)}$ are neighbourhoods of θ_0 and $\mathcal{N}_\lambda$ respectively;* **(vii)** *$\sup_{\lambda\in\Lambda_n}\sup_{\theta\in\Theta}\sup_{z\in\mathcal{Z}}|\rho_j(\lambda' g(z,\theta))| \leq C_{\rho_j}$ for $j = 1, 2$;* **(viii)** *there exists $C_{B,\mathcal{A}} > 0$ such that for all $k \in \{1, ..., d_g\}$, $\sup_{A\in\mathcal{A}}\left|\mathbb{E}[g^{(k)}(z,\theta_0)\ \mathbb{I}\{z \in A\}]\right| \leq C_{B,\mathcal{A}}$.*

Assumption 2.2 involves assumptions required for GEL estimation and boundedness conditions similar to those of Otsu [26] for large deviation analysis. The assumptions guarantee uniqueness of the GEL estimator and restrict the local curvature of the GEL objective function with respect to λ in a neighborhood of 0. Assumption 2.2(ii) relates to GEL estimation of the Lagrange multiplier (see Section 2.2 below) and is similar to the restriction on the set Λ_n defined in Lemma A1 of Newey and Smith ([25], p.239).

In Assumption 2.2(viii), the scalar $C_{B,\mathcal{A}}$ depends on the class of sets $\mathcal{A}$ considered and limits the average behaviour of individual sets $A \in \mathcal{A}$ (see Section 3.1 for examples of the class of sets $\mathcal{A}$). While C_g is a crude bound on the moment function, $C_{B,\mathcal{A}}$ limits the expected deviation of $\sup_{A\in\mathcal{A}}\left|\mathbb{E}[g^{(k)}(z,\theta_0)\mathbb{I}\{z \in \mathcal{A}\}]\right|$ for any moment function, $g^{(k)}$, $k \in \{1, ..., d_g\}$. If the moment function is relative stable across different subsets of the population, then $C_{B,\mathcal{A}}$ will be much smaller than C_g.

Assumption 2.3. *Let $\Omega_n(\theta) = \sum_{i=1}^n g(z_i,\theta)g(z_i,\theta)'/n$. For all n, there exist constants $C_\Omega^{(1)}$, $c_\Omega^{(1)} > 1$ such that $\mathbb{P}(\sup_{\theta\in\Theta}||\Omega_n^{-1}(\theta) - \Omega^{-1}(\theta)|| > \epsilon) \leq C_\Omega^{(1)}\mathbb{P}(\sup_{\theta\in\Theta}||\Omega_n(\theta) - \Omega(\theta)|| > \epsilon/c_\Omega^{(1)})$ for any $\epsilon > 0$.*

This assumption concerns the convergence, continuity and positive definiteness of covariance matrices. It is shown that there exist positive constants c_Ω and C_Ω such that $\mathbb{P}(\sup_{\theta\in\Theta}|\Omega_n(\theta)-\Omega(\theta)|>\epsilon)\leq C_\Omega\exp(-c_\Omega n)$. By noting $\Omega_n^{-1}(\theta)-\Omega^{-1}(\theta)=\Omega_n^{-1}(\theta)(\Omega(\theta)-\Omega_n(\theta))\Omega^{-1}(\theta)$, Assumption 2.1(iv) and the Cauchy-Schwarz inequality, Assumption 2.3 can be verified (cf. Otsu [26], discussion of Assumptions 2.3 and 2.4, p.324).

2.2 GEL estimation

The GEL estimator, Smith [33], is defined as follows. Let

$$\hat{P}_n(\theta,\lambda)=\frac{1}{n}\sum_{i=1}^{n}[\rho(\lambda' g(z_i,\theta))-\rho_0]$$

where the function $\rho(\cdot)$ satisfies Assumption 2.2(i). The GEL estimator of θ is defined as

$$\hat{\theta}=\underset{\theta\in\Theta}{\operatorname{argmin}}\sup_{\lambda\in\Lambda_n}\hat{P}_n(\theta,\lambda)$$

for the search set Λ_n defined in Assumption 2.2(ii). For any $\theta\in\Theta$, an estimator of the d_g-vector of auxiliary parameters λ is given by $\hat{\lambda}(\theta)=\operatorname{argmax}_{\lambda\in\Lambda_n}\hat{P}_n(\theta,\lambda(\theta))$. The first-order condition for λ imposes the sample moment constraint $\sum_{i=1}^n\hat{\pi}_i g(z_i,\hat{\theta})=0$ where the GEL implied probabilities are

$$\hat{\pi}_i=\frac{\rho_1(\hat{\lambda}' g(z_i,\hat{\theta}))}{\sum_{j=1}^n\rho_1(\hat{\lambda}' g(z_j,\hat{\theta}))},\ \ (i=1,...,n).$$

where $\hat{\lambda}=\hat{\lambda}(\hat{\theta})$.

A special case includes the continuously-updated GMM (CU-GMM, Hansen et al. [15]), where $\rho(v)=-\frac{1}{2}v^2-v$. CU-GMM estimation returns closed form solutions for the Lagrange multiplier $\hat{\lambda}(\theta)=-V_n(\theta)^{-1}g_n(\theta)$ where $g_n(\theta)=\sum_{i=1}^n g(z_i,\theta)/n$ and $V_n(\theta)=\sum_{i=1}^n g(z_i,\theta)(g(z_i,\theta)-g_n(\theta))'$, and implied probabilities

$$\hat{\pi}_i^{CU-GMM}=\frac{1}{n}-\frac{1}{n}g_n(\hat{\theta})'V_n(\hat{\theta})^{-1}(g(z_i,\hat{\theta})-g_n(\hat{\theta})),\ \ (i=1,...,n). \tag{2.2}$$

GEL methods have been shown to have finite-sample advantages over GMM. With evidence from Monte Carlo studies, Hansen et al. [15] and Imbens et al. [17] report that tests of overidentifying moment restrictions based on CU-GMM and particular GEL methods respectively, have attractive size properties relative to Hansen's [14] J test. For estimation, Newey et al. [24] present Monte Carlo results indicating smaller bias properties of GEL estimators in small samples compared to GMM. On a practical note, in contrast to two-step efficient GMM estimation, GEL methods do not require estimation of the Jacobian or covariance matrix Ω.

While the first-order asymptotic properties of GEL and GMM estimators are identical, Newey and Smith [25] show the second-order asymptotic bias for GEL estimators comprises of fewer components compared to that of GMM, with the empirical likelihood (EL) estimator being the best by this measure. However, EL methods may not be particularly robust to misspecification; the first-order conditions determining the EL estimator can be unstable resulting in poor behaviour. This is illustrated by Imbens et al. [17] and Schennach [32], where it is also shown that, in contrast, the exponential tilting estimator possesses attractive robustness properties under misspecification.

2.3 GEL implied probabilities under heterogeneity

A contrast between the GEL implied probabilities and the EDF weight $1/n$ provides potential information-theoretic inference on the validity of the moment restrictions (2.1). Intuitively implied probabilities reveal how much weight is being placed on an observation by GEL estimation. Under correct specification of moment restrictions, the following first-order Taylor expansion holds, see Lemma A1 of RS,

$$\hat{\pi}_i = \frac{1}{n} + \frac{1}{n} g(z_i, \hat{\theta})' \hat{\lambda}(1 + o_p(1)) + O_p(n^{-2}), \quad (i = 1, ..., n).$$

Under correct specification, it can be shown that $|g(z_i, \hat{\theta})' \hat{\lambda}|$ converges in probability to zero for every $i = 1, ..., n$, Lemma A1, Newey and Smith [25]. Therefore the weights closely approximate $1/n$ so that all observations are given approximately equal weight and the GEL-weighted moment function estimate $\sum_{i=1}^{n} \hat{\pi}_i g(z_i, \hat{\theta})$ is close to the empirical average $\sum_{i=1}^{n} g(z_i, \hat{\theta})/n$.

To illustrate these ideas, consider an example where the available data $\{z_i\}_{i=1}^{n}$ is assumed to follow a normal distribution with mean θ_1 and variance θ_2^2. Then the following moment conditions are valid

$$\begin{aligned} \mathbb{E}[z - \theta_1] &= 0 \\ \mathbb{E}[(z - \theta_1)^2 - \theta_2^2] &= 0 \\ \mathbb{E}[z^3 - \theta_1(\theta_1^2 + 3\theta_2^2)] &= 0. \end{aligned}$$

The moment conditions describe the mean, variance and skewness, respectively, of a normal distribution. For sample size n, the parameters for observations $\{z_i\}_{i=1}^{0.4n}$ and $\{z_i\}_{i=0.6n+1}^{n}$ are $\theta_1 = 1$ and $\theta_2 = 1$. For observations $\{z_i\}_{i=0.4n+1}^{0.6n}$ the parameters are $\theta_1 = 4$ and $\theta_2 = 1$. The GEL method employed is CU-GMM. Figure 2.1 shows how the implied probabilities $\hat{\pi}_i^{CU-GMM}$, $(i = 1, ..., n)$ from (2.2) vary with each observation for sample sizes $n = 100$, 250, 1000.

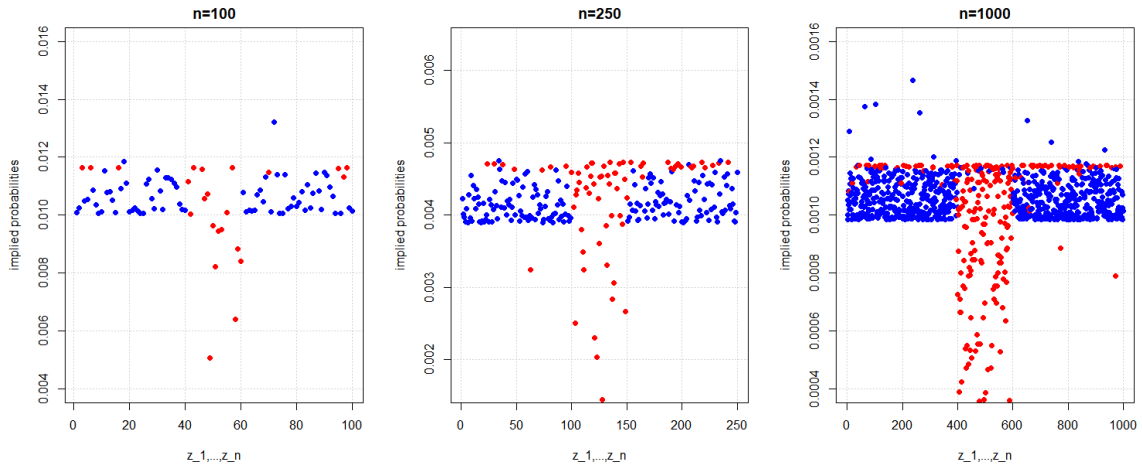


Figure 2.1. Simulated CU-GMM implied probabilities under parameter instability. The vertical axis shows the CU-GMM implied probability weights, and the horizontal axis shows the indices of the observations $\{z_i\}_{i=1}^n$.

Observations are grouped and colour-coded by k-means clustering with $k = 2$. The results show systematic deviations of the implied probabilities away from $1/n$. The patterns become more pronounced as the sample size increases.

Figure 2.1 shows even basic clustering algorithms can identify problem regions. Furthermore, in such simple cases of structural change, since these observations are ordered, patterns of parameter instability are easier to spot by inspection. In practice, possibly complicated non-linear functions of individual characteristics may be driving heterogeneous behaviour. This motivates the development of a general framework to find more subtle and complex patterns of misspecification that may exist in moment condition models.

3 Large deviations bounds under moment restrictions

3.1 Vapnik-Chervonenkis theory

Let $\mathcal{A}$ be a class of sets of the covariate space of z, i.e. $\mathcal{A} \subset \mathcal{Z}$. Consider probability bounds for the Kolmogorov-Smirnov distance between a GEL measure and an empirical measure taken over sets $A \in \mathcal{A}$ of the covariate space, that is,

$$\sup_{A\in\mathcal{A}} \left| \hat{F}(A) - F_n(A) \right| \tag{3.1}$$

where $\hat{F}(A) = \sum_{i=1}^n \hat{\pi}_i \mathbb{I}\{z_i \in A\}$ and $F_n(A) = \sum_{i=1}^n \mathbb{I}\{z_i \in A\}/n$.

Kolmogorov-Smirnov statistics concern a large literature in empirical process theory. If unrestricted, the class $\mathcal{A}$ is too large to derive meaningful bounds. If too restricted the class $\mathcal{A}$ will not be able to detect more complex misspecification patterns that may be related to

characteristics of z. A restricted class that is rich enough from which to draw meaningful conclusions is a VC class. VC theory introduces a measure of combinatorial richness of a class of sets $\mathcal{A}$. In statistical learning theory, the VC dimension for a data-generating process is a measure of sample complexity. The higher the VC dimension, the more complex a sample can be. The VC dimension is therefore a key factor in bounding the probability of observing an unrepresentative sample. The following definitions are as outlined in Section 12.4 of Devroye et al. [6]; also see Chapter 2.6 of van der Vaart and Wellner [35].

Definition (**Shattering coefficients**)**.** Let $\mathcal{A}$ be a collection of subsets from $\mathcal{Z}$. An arbitrary set of n points possesses 2^n subsets. For $(z_1, ..., z_n) \in \mathcal{Z}$, let $N_{\mathcal{A}}(z_1, ..., z_n)$ be the number of different sets in $\{\{z_1, ..., z_n\} \cap A : A \in \mathcal{A}\}$. The n-th shattering coefficient of $\mathcal{A}$ is $s(\mathcal{A}, n) = \max_{(z_1,...,z_n)\in\mathcal{Z}} N_{\mathcal{A}}(z_1, ..., z_n)$. That is, the shattering coefficient is the maximal number of different subsets of n points that can be picked out by the class of sets $\mathcal{A}$.

Definition (**VC dimension**)**.** Let $\mathcal{A}$ be a collection of subsets from $\mathcal{Z}$ with $|\mathcal{A}| \geq 2$. The largest integer $k \geq 1$ for which $s(\mathcal{A}, k) = 2^k$ is the Vapnik-Chervonenkis (VC) dimension of the class $\mathcal{A}$, and is denoted by v. Therefore, v is the maximal number of points in $\mathcal{Z}$ that can be shattered by $\mathcal{A}$. $\mathcal{A}$ is a VC class if $v < \infty$.

Therefore the following assumption is on the combinatorial richness of the class $\mathcal{A}$.

Assumption. 3.1. *$\mathcal{A}$ is a VC class of subsets of $\mathcal{Z}$ with VC dimension $v < \infty$.*

There are many ways of generating a collection of subsets of $\mathcal{Z}$ with a finite VC dimension. Consider the following examples of the class $\mathcal{A}$. Note that a set of only the blue points cannot be picked out by the class of classifiers considered.

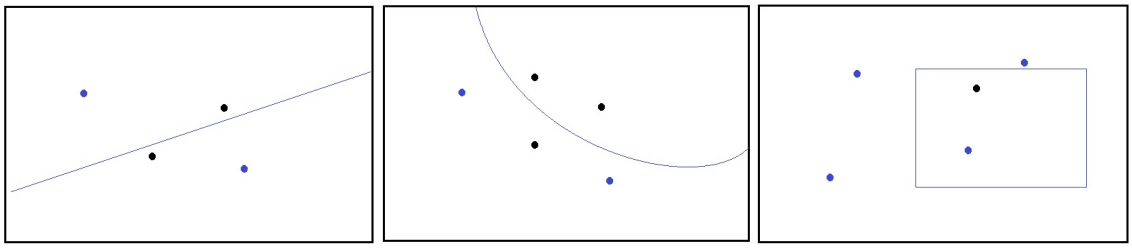


Figure 3.1. Examples of VC classes of sets.

1. *Linear discriminators*: $\mathcal{A}$ is the class of subsets of $\mathbb{R}^d$ of form $\{z : az + b \geq 0\}$. Then the VC dimension is $v = d + 1$. In $\mathbb{R}^2$, 3 points can be shattered but not 4.

2. *Quadratic discriminators:* $\mathcal{A}$ is the class of close balls in $\mathbb{R}^d$ of form $\{z : \sum_{i=1}^{d} |z^{(i)} - a_i|^2 \leq b\}$. Then the VC dimension is $v = d + 2$. In $\mathbb{R}^2$, 4 points can be shattered but not 5.

3. *Rectangles:* $\mathcal{A} = \{$class of all rectangles in $\mathbb{R}^d\}$, then the VC dimension is $v = 2d$. In $\mathbb{R}^2$, 4 points can be shattered but not 5.

The following result is a generalised version of the original VC inequality, Vapnik and Chervonenkis [36], for moment functions. Let $\mathbb{P}_n(.)$ denote the probability taken over different realisations of the random sample.

Lemma. 3.1. (VC Inequality for Moment Functions). *Under Assumptions 2.1-2.3 and 3.1, for all $\epsilon > 0$ and $n > \max\{2v+1, 8C_g^2/\epsilon^2\}$, for any moment function $g^{(k)}$, $k \in \{1, ..., d_g\}$,*

$$\mathbb{P}_n\Big(\sup_{A\in\mathcal{A}} \Big|\frac{1}{n}\sum_{i=1}^{n} g^{(k)}(z_i,\theta_0)\mathbb{I}\{z_i \in A\} - \mathbb{E}[g^{(k)}(z,\theta_0)\mathbb{I}\{z \in A\}]\Big| > \epsilon\Big) \leq 8\Big(\frac{16enC_g}{(2v+1)\epsilon}\Big)^{2v+1} \exp\Big(-\frac{\epsilon^2 n}{128C_g^2}\Big).$$

The proof of this result is given for Lemma C.5 in the Appendix. The rate of convergence depends on the range of values the moment function can produce. Furthermore, the bound is also influenced by the VC dimension of the class of sets $\mathcal{A}$; if the class of sets considered has a low VC dimension, the bound will be tighter. In practice, θ_0 is unknown. However, for the case of the continuously-updating GMM estimator, a non-asymptotic tail bound for deviations $||\hat{\theta} - \theta_0||$ are derived in Theorem 3.2. Therefore by the triangle and Cauchy-Schwarz inequalities, using Lemma 3.1 and Theorem 3.2 below provides a non-asymptotic tail bound for deviations $\sup_{A\in\mathcal{A}} |n^{-1}\sum_{i=1}^{n} g^{(k)}(z_i,\hat{\theta})\mathbb{I}\{z_i \in A\} - \mathbb{E}[g^{(k)}(z,\theta_0)\mathbb{I}\{z \in A\}]|$, $(k = 1, ..., d_g)$.

3.2 Large deviations of GEL distribution functions

To bound (3.1), to derive results for the general GEL class we consider a first-order Taylor approximation so that asymptotic GEL results hold. In particular, replace $\{\hat{\pi}_i\}_{i=1}^n$ by $\{\tilde{\pi}_i\}_{i=1}^n$ where

$$\tilde{\pi}_i \;\; = \;\; n^{-1} + n^{-1}\hat{\lambda}' g(z_i, \hat{\theta}),$$

with $\hat{\theta}$ and $\hat{\lambda}$ the GEL estimators of θ_0 and Lagrange multiplier λ respectively. The GEL estimator of the distribution function of z is thus $\tilde{F}(A) = \sum_{i=1}^n \tilde{\pi}_i \mathbb{I}\{z_i \in A\}$. It can be shown that $\hat{F}(A) = \tilde{F}(A) + O_p(n^{-1})$ where $\hat{F}(A) = \hat{\pi}_i \mathbb{I}\{z_i \in A\}$, and the following result from RS holds,

$$\sqrt{n}(\tilde{F}(A) - F_n(A)) \xrightarrow{d} \mathcal{N}\big(0, V(A)\big) \tag{3.2}$$

where $V(A) = B(A)'PB(A)$ with $B(A) = \mathbb{E}[g(z,\theta_0)\mathbb{I}\{z \in A\}]$ and $P = \Omega^{-1} - \Omega^{-1}G(G'\Omega^{-1}G)^{-1}G'\Omega^{-1}$, where Ω and G are defined in Assumption 2.1.

Consider choices of $A \in \mathcal{A}$ such that the difference $\tilde{F}(A) - F_n(A)$ is maximised.

Theorem 3.1. (Large Deviation Bounds of GEL Distribution Functions). *Under Assumptions 2.1-2.3 and 3.1, there exist positive constants C_{gel} and c_{gel} such that for any $\epsilon > 0$,*

$$\mathbb{P}_n\Big(\sup_{A\in\mathcal{A}} \big|\tilde{F}_n(A) - F_n(A)\big| > \epsilon\Big) \leq C_{gel} \exp\{-c_{gel} n\}.$$

REMARK 3.1A. By the traditional VC inequality, the error probability of $\sup_{A\in\mathcal{A}}\left|F_n(A) - F(A)\right|$ is exponentially small, where $F(A) = \mathbb{P}(z \in A)$ is the true distribution function of z (Vapnik and Chervonenkis [36]). Therefore, by the triangle inequality, Theorem 3.1 also establishes the estimation error probability of the GEL distribution function, $\sup_{A\in\mathcal{A}}\left|\tilde{F}_n(A) - F(A)\right|$ convergences to zero at an exponentially fast rate.

REMARK 3.1B. Theorem 3.1 is analogous to the large deviation bound for GEL estimators $\mathbb{P}_n(||\hat{\theta} - \theta_0|| > \epsilon) \leq C_\theta \exp\{-c_\theta n\}$ of Otsu [26], see Lemma B.5 in the Appendix, however Otsu derives bounds under more generality whereby local misspecification is permitted. The result can also be used for more detailed estimation error analysis.

3.3 Non-asymptotic bounds for continuously-updating GMM estimation

CU-GMM is a popular estimator within the GEL class. Antoine et al. [3] discuss advantages of CU-GMM relative to two-step GMM and other GEL estimators. Notably, the first step maximisation over $\lambda \in \Lambda_n$, see Section 2.2, returns a closed form solution. The CU-GMM criteron is $Q_n(\theta) = g_n(\theta)'\Omega_n^{-1}(\theta)g_n(\theta)$ where $g_n(\theta)$ is defined in Section 2.2, and $\Omega_n(\theta) = \sum_{i=1}^n g(z_i,\theta)g(z_i,\theta)'/n$. The population counterpart of $Q_n(\theta)$ is $Q_0(\theta) = \mathbb{E}[g(z,\theta)]'\Omega(\theta)^{-1}\mathbb{E}[g(z,\theta)]$. The CU-GMM estimator is $\hat{\theta} = \arg\min_{\theta\in\Theta} Q_n(\theta)$.

Theorem 3.2. (Large Deviation Bounds for CU-GMM Estimation). *Under Assumptions 2.1-2.3 and 3.1, if Θ is a ball in $\mathbb{R}^{d_\theta}$ of radius $R > 0$, the CU-GMM estimator satisfies*

$$\mathbb{P}_n\Big(||\hat{\theta} - \theta_0|| > \epsilon\Big) \leq \omega_1 \exp\{-\omega_2\delta^2 n\} + \omega_3 \exp\Big\{-\omega_4 n\Big\}$$

for positive constants

$$\begin{aligned}
\omega_1 &= 3\max\Big\{8d_g\Big(\frac{48d_gC_GR}{\kappa_1}\Big)^{d_\theta}, 8d_g\Big(\frac{48d_gC_GR}{\kappa_2}\Big)^{d_\theta}, 8C_\Omega^{(1)}d_g^2\Big(\frac{96d_gc_\Omega^{(1)}C_gC_GR}{\kappa_3}\Big)^{d_\theta}\Big\} \\
\omega_2 &= \min\Big\{\frac{\bar{\kappa}_1^2}{128d_g^2C_g^2}, \frac{\bar{\kappa}_2^2}{128d_g^2C_g^2}, \frac{\bar{\kappa}_3^2}{128(c_\Omega^{(1)})^2d_g^2C_g^4}\Big\} \\
\omega_3 &= 8C_\Omega^{(1)}d_g^2(96d_gc_\Omega^{(1)}C_gC_GR)^p \\
\omega_4 &= \frac{1}{128(c_\Omega^{(1)})^2d_g^2C_g^4}
\end{aligned}$$

and $\delta = \delta(\epsilon) = \inf_{\theta\in\Theta,|\theta-\theta_0|>\epsilon} Q_0(\theta) - Q_0(\theta_0) > 0$, where $\kappa_1 = \delta\delta_{\Omega,\min}/(9d_gC_g(1+\delta_{\Omega,\min}))$, $\kappa_2 = \delta/(9d_gC_g\delta_{\Omega,\min})$, $\kappa_3 = \delta/(9d_g^2C_g^2)$, $\bar{\kappa}_1 = \delta_{\Omega,\min}/(9d_gC_g(1+\delta_{\Omega,\min}))$, $\bar{\kappa}_2 = 1/(9d_gC_g\delta_{\Omega,\min})$ and $\bar{\kappa}_3 = 1/(9d_g^2C_g^2)$.

REMARK 3.2A. Similarly to Otsu [26], the CU-GMM estimator has exponentially small deviation error probability. In contrast, however, explicit constants for the bound are derived that characterise the large deviation behaviour.

REMARK 3.2B. Theorem 3.2 shows that the probability bound depends on the strength of identification $\delta = \delta(\epsilon) > 0$. Since the true value θ_0 minimises $Q_0(\theta)$, $\inf_{\theta\in\Theta,\,|\theta-\theta_0|>\epsilon} Q_0(\theta) - Q_0(\theta_0) > 0$. If θ_0 is strongly identified, then $\delta(\epsilon) = \inf_{\theta\in\Theta,\,|\theta-\theta_0|>\epsilon} Q_0(\theta) - Q_0(\theta_0)$ is large for any $\epsilon > 0$ resulting in a stronger bound for larger given $\epsilon > 0$.

REMARK 3.2C. The constants ω_1 and ω_3 depend on the size of the search for θ_0. In particular, in order to apply covering arguments, Θ is a ball in $\mathbb{R}^p$ of radius $R > 0$. The larger the search, the weaker the bound.

Theorem 3.2 enables non-asymptotic bounds analogous to Theorem 3.1 to be derived for CU-GMM.

Theorem 3.3. (Large Deviation Bound for CU-GMM Distribution Functions). *Let* $\tilde{F}_n(A) - F_n(A) = f_n(\hat{\theta}, A) = g_n(\hat{\theta})\Omega_n^{-1}(\hat{\theta})\big(\sum_{i=1}^n g(z_i,\hat{\theta})\mathbb{I}\{z_i \in A\}/n\big)$. *Under Assumptions 2.1-2.3 and 3.1, if* Θ *is a ball in* $\mathbb{R}^{d_\theta}$ *of radius* $R > 0$, *for CU-GMM, for any* $\epsilon > 0$,

$$\begin{aligned}\mathbb{P}_n(|f_n(\hat{\theta},A)| > \epsilon) \;\leq\; & \min\Bigg\{\Big[\omega_1\exp\{-\omega_2\delta^2(\epsilon^\star)n\} + \omega_3\exp\Big\{-\omega_4 n\Big\}\Big] + \Big[2d_g\exp\Big(\frac{-n(\epsilon^{\star\star})^2}{2d_g^2C_g^2}\Big)\Big] \\ & , \Big[\omega_1\exp\{-\omega_2\delta^2(\epsilon_a)n\} + \omega_3\exp\Big\{-\omega_4 n\Big\}\Big] + \Big[8d_g\Big(\frac{16end_gC_g}{(2v+1)\epsilon_b}\Big)^{2v+1}\exp\Big(-\frac{\epsilon_b^2 n}{128C_g^2d_g^2}\Big)\Big] \\ & + \Big[\omega_1\exp\{-\omega_2\delta^2(\epsilon_c)n\} + \omega_3\exp\Big\{-\omega_4 n\Big\}\Big] + \Big[2d_g\exp\Big(\frac{-n\epsilon_d^2}{2d_g^2C_g^2}\Big)\Big]\Bigg\} + \omega_3\exp\Big\{-\omega_4 n\Big\}\end{aligned}$$

where the constants ω_1, ω_2, ω_3 *and* ω_4 *are defined in Theorem 3.2,* $\epsilon^\star = \epsilon\delta_{\Omega,\min}/(2(1+\delta_{\Omega,\min})d_gC_g)$, $\epsilon^{\star\star} = \epsilon\delta_{\Omega,\min}/(2(1+\delta_{\Omega,\min})d_gC_g)$, $\epsilon_a = \epsilon\delta_{\Omega,\min}/(4d_g^2C_GC_g(1+\delta_{\Omega,\min}))$, $\epsilon_b = \epsilon\delta_{\Omega,\min}/(4d_gC_g(1+\delta_{\Omega,\min}))$, $\epsilon_c = \epsilon\delta_{\Omega,\min}/(4d_g^2C_GC_{B,\mathcal{A}}(1+\delta_{\Omega,\min}))$ *and* $\epsilon_d = \epsilon\delta_{\Omega,\min}/(4d_gC_{B,\mathcal{A}}(1+\delta_{\Omega,\min}))$.

REMARK 3.3A. These bounds are based on a first-order approximation of CU-GMM implied probabilities. Since these implied probabilities have a closed form solution, it is possible to derive similar bounds for CU-GMM using the true implied probabilities instead of using a first-order approximation.

REMARK 3.3B. As in Theorem 3.2, the strength of identification δ influences the quality of the bound.

4 Unrepresentative region estimation

Testing the null hypothesis H_0 : there exists $\theta_0 \in \Theta$ such that $\mathbb{E}[g(z,\theta_0)] = 0$, versus the general alternative H_1: there exists no $\theta \in \Theta$ such that $\mathbb{E}[g(z,\theta)] \neq 0$, may be difficult when the moment condition model is violated for a small subgroup of the population. In general, the usual tests for overidentifying moment restrictions may not be able to reject the validty of the moment conditions since they are largely true when averaged over the population.

Yet it is often of interest to identify those parts of the covariate space for which behaviour characterised by a moment condition model is different from the average. For example, if the moment indicator $g(z, \theta_0)$ describes the average treatment effect, see, for example, Hahn [11], analysis of $\mathbb{E}[g(z, \theta_0)\mathbb{I}\{z \in A\}]$ is important for identifying subsets of the population that are particularly responsive to the treatment. Using this idea, Kitagawa and Tetenov [19] find treatment rules that maximise a welfare function relating to the average treatment effect in particular regions. The class of sets considered is a VC class; for example, a class of linear eligibility rules can be described $\mathcal{A} = \{z \in \mathbb{R}^{d_z} : \alpha + z'\beta \geq 0, \alpha \in \mathbb{R}, \beta \in \mathbb{R}^{d_z}\}$. This principle is generalised here for inference in moment condition models. We introduce the notion of the *unrepresentative region*: the covariate region $A \in \mathcal{A} \subset \mathcal{Z}$ for which $\mathbb{E}[g(z, \theta_0)\mathbb{I}\{z \in A\}]$ is significantly different from zero.

4.1 Unconditional and conditional moment restrictions

It is important to note that the unconditional moment condition model (2.1) does not imply that $\mathbb{E}[g^{(k)}(z, \theta_0)\mathbb{I}\{z \in A\}] = 0$ for any $k = 1, ..., d_g$, however, if $\mathbb{E}[g^{(k)}(z, \theta_0)\mathbb{I}\{z \in A\}] \approx 0$, it follows that the model expressed through the k-th moment restriction is representative for individuals with covariates in region $A \in \mathcal{A}$.

Suppose now that the vector of variables z consists of an outcome variable y and instruments x, such that the instrumental variables model $\mathbb{E}[g(z, \theta_0)|x] = 0$ holds. Then for any fixed region A in the space of instruments, the conditional moment restriction implies $\mathbb{E}[g(z, \theta_0)\mathbb{I}\{x \in A\}] = 0$ holds. Therefore, the notion of an unrepresentative region cannot apply for conditional moment restrictions if those regions are defined solely as sets of instrumental variables.

On the other hand, if the regions are defined in terms of the joint distribution of x and y, then in general $\mathbb{E}[g(z, \theta_0)\mathbb{I}\{z \in A\}] \neq 0$. One example of this are data-driven selection of regions, where an estimated region $\hat{A}$ is defined by a criterion involving x and y, and thus $\hat{A} = \hat{A}(z)$ can be considered a function of both x and y. We show next that analysis of such data-driven regions may shed light on possible neglected heterogeneity in the model.

4.2 Unrepresentative region

We first must define a general class of sets $\mathcal{A}$ of the space of w, where $w \subseteq z$. $\mathcal{A}$ must be a VC class; some examples are given in Section 3.1. In order to define the unrepresentative region the following additional quantities are introduced. For any set $A \in \mathcal{A}$, let $B^{(k)}(A) = \mathbb{E}[g^{(k)}(z, \theta_0)\mathbb{I}\{w \in A\}]$, $B_n^{(k)}(A) = \sum_{i=1}^n g^{(k)}(z_i, \theta_0)\mathbb{I}\{w \in A\}/n$; $\hat{B}_n^{(k)}(A) = \sum_{i=1}^n g^{(k)}(z_i, \hat{\theta})\mathbb{I}\{w \in A\}/n$, $(k = 1, ..., d_g)$; $W(A) = \sum_{k=1}^{d_g}(B^{(k)}(A))^2$; $W_n(A) = \sum_{k=1}^{d_g}(B_n^{(k)}(A))^2$; $\hat{W}_n(A) = \sum_{k=1}^{d_g}(\hat{B}_n^{(k)}(A))^2$.

Definition (unrepresentative region). The unrepresentative region (UR) in $\mathcal{A}$ with respect to the moment condition $\mathbb{E}[g(z, \theta_0)] = 0$ is defined to be $\tilde{A} = \arg\max_{A \in \mathcal{A}} W(A)$.

An estimate $\hat{A}$ of the UR $\tilde{A}$ is obtained by maximising its sample analogue evaluated at the GEL estimator $\hat{\theta}$, that is,

$$\hat{A} = \arg\max_{A \in \mathcal{A}} \hat{W}_n(A). \tag{4.1}$$

The consistency concept employed here is in terms of the Hausdorff metric. For the set $\tilde{A}$, for any w, w' and w'', let $d(w, \tilde{A}) = \inf_{w' \in \tilde{A}} ||w - w'||$ and $d(w, \hat{A}) = \inf_{w'' \in \hat{A}} ||w - w''||$. The Hausdorff distance between two sets $\hat{A}$ and $\tilde{A}$ defined as $d_H(\hat{A}, \tilde{A}) = \max\big\{ \sup_{w \in \hat{A}} d(w, \tilde{A}), \sup_{w \in \tilde{A}} d(w, \hat{A}) \big\}$. The estimated UR $\hat{A}$ is said to be a consistent estimator for the true UR $\tilde{A}$ if $d_H(\hat{A}, \tilde{A}) = o_p(1)$.

Theorem 4.1. *Under Assumptions 2.1-2.3 and 3.1, if a unique UR $\tilde{A}$ exists,* **(i)** $W(\tilde{A}) - W(\hat{A}) = o_p(1)$**; (ii)** $d_H(\hat{A}, \tilde{A}) = o_p(1)$.

Remark 4.1a. In order to ensure existence of an UR, further restrictions may need to be imposed on the class of sets $\mathcal{A}$. At least one set A in the class $\mathcal{A}$ must not be so large that the moment condition $\mathbb{E}[g(z, \theta_0)] = 0$ implies $\mathbb{E}[g(z, \theta_0)\mathbb{I}\{w \in A\}] = 0$. On the other hand, A must not be so small that that the probability that $\mathbb{P}(w \in A) = 0$.

Remark 4.1b. Given the definition of the UR above, the size of sets in the class $\mathcal{A}$ should be further penalised to improve the ability of an UR to shed light on subtle forms of neglected heterogeneity. For example, the simulation studies in Sections 4.3 and 5.1 partition the space of w into equal-spaced cubes in $\mathbb{R}^{\dim(w)}$.

4.3 Simulation study

This simulation experiment examines the validity of a simple instrumental variables model with binary instruments. Let $w = (w_1, w_2, w_3)'$ be a vector of independent instruments taking values of $\{0, 1, 2\}$, each binomially distributed $B(2, 0.3)$. x is a single endogenous variable and y is a single outcome variable. The data are generated from

$$\begin{aligned} x_i &= \frac{1}{3}\Big(\sum_{j=1}^{3} w_{ji}\Big) + v_i \\ y_i &= \theta_0 x_i + u_i, \end{aligned}$$

where the errors have the form $v_i = \epsilon_i + e_{xi}$, $u_i = \epsilon_i + e_{yi}$ and $\epsilon_i, e_{xi}, e_{yi} \sim N(0, 1)$, and w_{ji} is the i-th observation of w_j $(j = 1, 2, 3)$. The true value of causal effect is set at $\theta_0 = 0.1$.

This is the model studied in Davies et al. [5] and is commonly employed in Mendelian randomisation studies to investigate the relationship between modifiable exposures and diseases. For example, w is a vector of genetic markers with its values representing allele frequencies, x is a continuous exposure, and y is a trait or outcome variable. For a measure of the strength

of instruments, we note that the concentration parameter for our model is given by $0.07n$ (see Davies et al. [5], equation 14, p.458).

θ_0 is estimated by CU-GMM. Let $g_n(\theta) = \sum_{i=1}^n w_i(y_i - \theta x_i)/n$ and $\Omega_n(\theta) = \sum_{i=1}^n w_i w_i'(y_i - \theta x_i)^2/n$, where $w_i = (w_{1i}, w_{2i}, w_{3i})'$, $(i = 1, ..., n)$. The CU-GMM estimator (CUE) is as described in Section 3.3.

The covariate space of instruments is given by $\mathcal{A} = \{0, 1, 2\}^3$ resulting in 27 distinct possible values of the vector w. $\mathcal{A}$ can be pictured as a 3 by 3 cubic with each cube representing combinations $\{w_1 = a, w_2 = b, w_3 = c | a, b, c \in \{0, 1, 2\}\}$. In terms of VC theory, $\mathcal{A}$ is a special case of the class of rectangles in $\mathbb{R}^3$ and thus $\mathcal{A}$ has a VC dimension of 6.

Consider a situation of parameter heterogeneity such that there is an extra effect for those individuals with characteristics $\{w_1 = 1, w_2 = 1, w_3 = 0\}$. Given the binomial distribution of w_j $(j = 1, 2, 3)$, individuals with these characteristics constitute approximately 8.6% of the sample. This is the UR $\tilde{A}$. In order to estimate this region, we define the function $W_n(A) = \sum_{j=1}^3 (\sum_{i=1}^n w_{ji}(y_i - \hat{\theta} x_i)\mathbb{I}\{w_i \in A\}/n)^2$, and calculate $\hat{A} = \arg\min_{A \in \mathcal{A}} W_n(A)$.

Under such parameter heterogeneity, $y = (\theta_0 + \kappa \mathbb{I}\{w_1 = 1, w_2 = 1, w_3 = 0\})x + u$, where κ is a parameter that represents the extent of the heterogeneity problem. If $\kappa = 0$, there is no heterogeneity and the instrumental variables model is correctly specified. We proceed with estimation neglecting the potential heterogeneity problem, that is, assuming $\kappa = 0$.

The experiments note the rate at which the estimated unpresentative region $\hat{A}$ picks out the set $\tilde{A}$ for varying values of κ and sample size n. The results reported are averaged over 5000 simulations for each experiment.

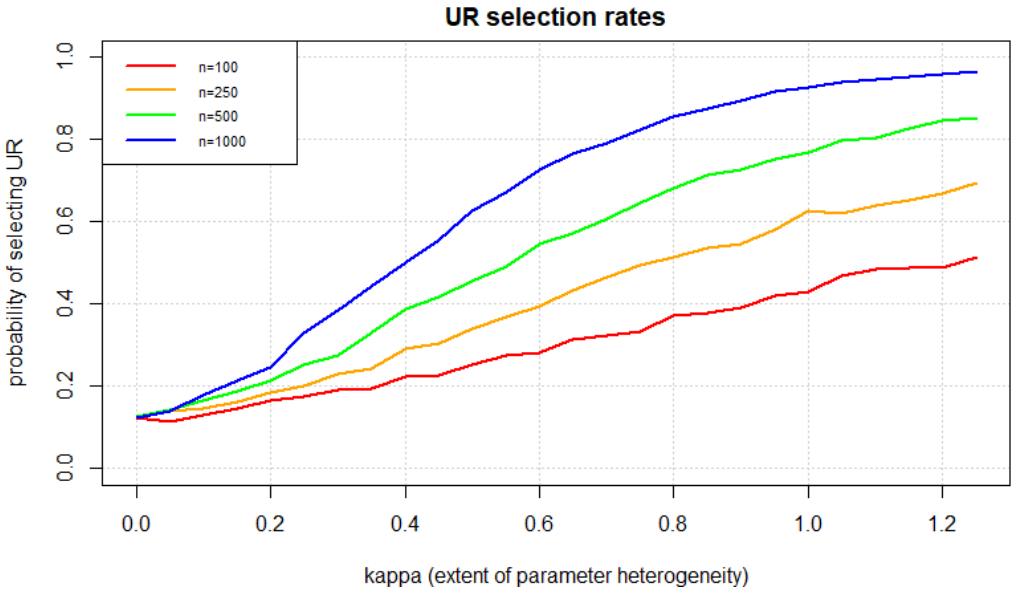


Figure 4.1. Rate at which $\hat{A} = \tilde{A}$ for sample sizes $n = 100, 250, 500, 1000$.

Figure 4.1 illustrates that the ability to pick out the UR depends heavily on the sample size

and the extent of the heterogeneity problem. For higher values of κ, or larger sample sizes, the UR is selected at a higher rate.

Interestingly, in absence of model misspecification $\kappa = 0$, the rate of selecting the UR is around 0.125. This suggests that, given the partition of the space of instruments into 27 sets (or cubes), all sets $A \in \mathcal{A}$ do not have an equal chance of selection as the UR under correct model specification.

5 Goodness-of-fit tests with data-driven partitions

RS propose goodness-of-fit test statistics based on a contrast of GEL implied probabilities with the corresponding EDF weights. Such statistics converge in distribution to a chi-squared random variable with degrees of freedom equal to the number of overidentifying moments.

This section revisits goodness-of-fit test statistics of RS but with data-driven partitions. Some restrictions need to be placed on the class of partitions considered to ensure the validity of asymptotic approximations employed for deriving the distribution of the test statistic under (2.1). Firstly, it is required that the number of separate sets s of the covariate space to be included in the test are at least equal to d_g (cf. the order condition, RS, p.9). Secondly, for sets $\{A_1, ..., A_s\}$ of z, the $d_g \times s$ matrix $B = \big(\mathbb{E}[g(z,\theta_0)\mathbb{I}\{z \in A_1\}], ..., \mathbb{E}[g(z,\theta_0)\mathbb{I}\{z \in A_s\}]\big)$ must have full row rank.

In order to ensure this, the class of partitions $\mathcal{A}$ of the covariate space may need to be curated depending on the application. A simple criteria is, for a given s, the class $\mathcal{A}$ may only consist of combinations of s sets $\{A_1, ...A_s\}$ such that: (i) B has full row rank; (ii) $\mathcal{A}$ has a finite VC dimension.

For construction of the test, we can select partitions according to the following sequential procedure. First, let the UR $\tilde{A}_1$ be such that

$$\tilde{A}_1 = \underset{A \in \mathcal{A}}{\arg\max}\ W(A).$$

Then, for $j = 2, ..., s$, define

$$\tilde{A}_j = \underset{A \in \mathcal{A}\backslash\{\tilde{A}_1, ..., \tilde{A}_{j-1}\}}{\arg\max}\ W(A)$$

where $\mathcal{A}\backslash\{\tilde{A}_1, ..., \tilde{A}_{j-1}\}$ is the class of sets $\mathcal{A}$ excluding $\tilde{A}_1, ..., \tilde{A}_{j-1}$. Furthermore, $\{\tilde{A}_1, ..., \tilde{A}_s\}$ must also be such that the $d_g \times s$ matrix $\big(\mathbb{E}[g(z,\theta_0)\mathbb{I}\{z \in \tilde{A}_1\}], ..., \mathbb{E}[g(z,\theta_0)\mathbb{I}\{z \in \tilde{A}_s\}]\big)$ has full row rank.

As in Section 4.2, we can estimate these sets by estimating the UR

$$\hat{A}_1 = \underset{A \in \mathcal{A}}{\arg\max}\ \hat{W}_n(A),$$

and for $j = 2, ..., s$, calculating

$$\hat{A}_j = \underset{A \in \mathcal{A} \setminus \{\hat{A}_1, ..., \hat{A}_{j-1}\}}{\arg \max} \hat{W}_n(A).$$

If $\mathcal{A}$ is a class of sets with VC dimension v, then for any $j = 2, ..., s$, $\mathcal{A}\setminus\{\tilde{A}_1, ..., \tilde{A}_{j-1}\}$ and $\mathcal{A}\setminus\{\hat{A}_1, ..., \hat{A}_{j-1}\}$ are classes of sets with VC dimension at most v. Therefore by applying similar arguments to Theorem 4.1, it can be shown that $d_H(\hat{A}_j, \tilde{A}_j) \xrightarrow{p} 0$, for any $j = 1, ..., s$.

Given sets $\{\hat{A}_1, ..., \hat{A}_s\}$ a goodness-of-fit test statistic can be constructed in the usual way by following Section 3.2 of RS. Let $\hat{B}_{s,n} = (\hat{B}_n(\hat{A}_1), ..., \hat{B}_n(\hat{A}_s))$ where $\hat{B}_n(A) = \sum_{i=1}^n g(z_i, \hat{\theta})\mathbb{I}\{z_i \in A\}/n$. For the implied probabilities from CU-GMM estimation (2.2) let $\hat{\mu}_n(A) = \sum_{i=1}^n (\hat{\pi}_i - (1/n))\mathbb{I}\{z_i \in A\}$. Construct the s-dimensional vector $\hat{\mu}_{s,n} = (\hat{\mu}_n(\hat{A}_1), ..., \hat{\mu}_n(\hat{A}_s))'$. The goodness-of-fit statistic is then defined as

$$P_{alt} = n\hat{\mu}'_{s,n}\hat{B}'_{s,n}(\hat{B}_{s,n}\hat{B}'_{s,n})^{-1}\hat{\Omega}(\hat{B}_{s,n}\hat{B}'_{s,n})^{-1}\hat{B}_{s,n}\hat{\mu}_{s,n},$$

where $\hat{\Omega} = \sum_{i=1}^n g(z_i, \hat{\theta})g(z_i, \hat{\theta})'/n$.

Corollary 5.1. *Under Assumptions 2.1-2.3 and 3.1, if unrepresentative regions $\{\tilde{A}_j\}_{j=1}^s$ exist, $P_{alt} \xrightarrow{d} \chi^2_{d_g - d_\theta}$.*

That is, the P_{alt} has the same null hypothesis asymptotic distribution as the RS test statistics and, thus, that of other commonly used GMM and GEL-based statistics used for testing overidentifying restrictions.

5.1 Simulation study

This section illustrates the use of UR calculations for goodness-of-fit testing with data-driven partitions. We consider a simulation study based on the asset pricing model studied by Imbens et al. [17] and RS. Let $z = (z_1, z_2)$ where z_1 and z_2 are generated independently from a $N(0, 0.16)$ distribution. Although the theoretical results obtained in previous sections use bounded moment functions, the assumptions could be weakened to requiring boundedness conditions in expectation following similar arguments to Otsu [26]. However, for this simulation study, we allow variables which, although having very small variance, are unbounded. The moment indicators

$$g(z, \theta) = \begin{pmatrix} \exp(-0.72 - \theta(z_1 + z_2) + 3z_2) - 1 \\ z_2(\exp(-0.72 - \theta(z_1 + z_2) + 3z_2) - 1) \end{pmatrix}$$

describe the preference parameter θ (with true value $\theta_0 = 3$), for a constant relative risk aversion utility function. See Gregory et al. [9], pp.218-219, for a full derivation of the moment indicators.

As in Section 4.3, we estimate θ_0 by the CUE $\hat{\theta}$. Our interest is in testing the null hypothesis $H_0 : \mathbb{E}[g(z, \theta_0)] = 0$ for some $\theta_0 \in \Theta$, versus the alternative $H_1 : \mathbb{E}[g(z, \theta)] \neq 0$ for all $\theta \in \Theta$.

5.1.1 Data-driven goodness-of-fit statistic

To construct the P_{alt} test, we need to create partitions. We partition the covariate space into equal-sized squares in $\mathbb{R}^2$ according to the quantiles of a $N(0, 0.16)$ distribution.

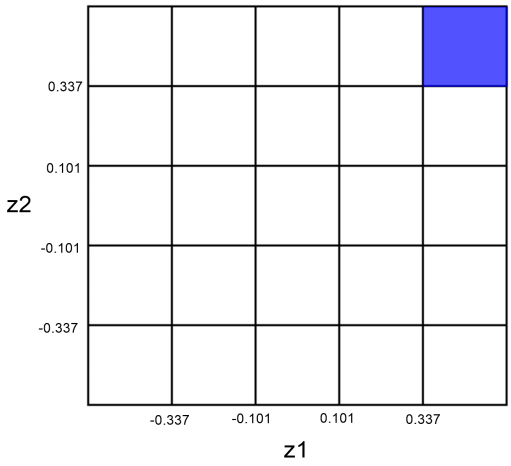


Figure 5.1. 5 by 5 partition of the covariate space into equal-spaced squares.

We consider partitions into 5 by 5, 6 by 6, and 7 by 7 squares, resulting in a total of $25 + 36 + 49 = 110$ regions. Denote the collection of these 110 sets by $\mathcal{A}$; the class of squares in $\mathbb{R}^2$ has VC dimension 4.

In order to construct the test statistic P_{alt}, we first estimate the set $\tilde{A} \in \mathcal{A}$ corresponding to the UR by $\hat{A} \in \mathcal{A}$. If $\hat{A}$ is one of the squares in the 5 by 5 partition, the P_{alt} statistic uses the 25 sets corresponding to the 5 by 5 to construct the test. For example, if those sets are denoted $\{A_1, ..., A_{25}\}$, then $\hat{B}_{25,n}$ is the 2×25 matrix with column $j \in \{1, ..., 25\}$ equal to $\sum_{i=1}^n g(z_i, \hat{\theta}) \mathbb{I}\{z_i \in A_j\}/n$, $\hat{\mu}_{25,n}$ is the 25-dimensional vector with its j-th element equal to $\sum_{i=1}^n (\hat{\pi}_i - (1/n)) \mathbb{I}\{z_i \in A_j\}$, so that $P_{alt} = n\hat{\mu}'_{25,n} \hat{B}'_{25,n} (\hat{B}_{25,n} \hat{B}'_{25,n})^{-1} \hat{\Omega} (\hat{B}_{25,n} \hat{B}'_{25,n})^{-1} \hat{B}_{25,n} \hat{\mu}_{25,n}$.

Similarly, if $\hat{A}$ is one of the squares in the 6 by 6 partition, the P_{alt} statistic uses the 36 sets corresponding to the 6 by 6 partition, and likewise the P_{alt} uses the 49 sets corresponding to the 7 by 7 partition if one of those squares in the 7 by 7 partition is $\hat{A}$. Furthermore, following RS, for calculating P_{alt} we employ the robust covariance estimator of Ω_0, $\hat{\Omega} = \big(\sum_{i=1}^n \hat{\pi}_i g(z_i, \hat{\theta}) g(z_i, \hat{\theta})'\big) \big(n \sum_{i=1}^n \hat{\pi}_i^2 g(z_i, \hat{\theta}) g(z_i, \hat{\theta})'\big)^{-1} \big(\sum_{i=1}^n \hat{\pi}_i g(z_i, \hat{\theta}) g(z_i, \hat{\theta})'\big)$.

5.1.2 Other test statistics

In order to evaluate the performance of the data-driven goodness-of-fit test, we compare its performance with other GEL based tests proposed in Smith [33] and RS. Again the particular

type of GEL method employed is CU-GMM.

Let $P_n(\hat{\theta}, \hat{\lambda}) = \sum_{i=1}^{n} \rho(\hat{\lambda}' g(z_i, \hat{\theta}))/n$, where $\rho(v) = -0.5v^2 - v$ for any $v \in \mathbb{R}$. The likelihood ratio (LR) statistic is given by $2n(P_n(\hat{\theta}, \hat{\lambda}) - \rho(0))$. The Lagrange multiplier (LM) statistic is given by $n\hat{\lambda}'\hat{\Omega}\hat{\lambda}$. The score statistic is given by $n g_n(\hat{\theta})'\hat{\Omega}_n(\hat{\theta})^{-1} g_n(\hat{\theta})$. Finally, RS also introduce two test statistics based on implied probabilities $P_a = \sum_{i=1}^{n}(n\hat{\pi}_i - 1)^2$ and $P_b = \sum_{i=1}^{n}((n\hat{\pi}_i - 1)^2/n\hat{\pi}_i)$.

For an α-level test we compare the value of these statistics with the $(1-\alpha)^{th}$ quantile of a chi-squared random variable with degrees of freedom 1, $\chi^2_{(1-\alpha),1}$.

5.1.3 Size analysis

Here we examine the finite-sample properties of the goodness-of-fit test statistic under correct model specification. We consider sample sizes of $n = 100, 250, 500$ and 1000 observations, with each experiment being replicated 10000 times. Table 5.1 reports the estimated size at nominal size levels 0.01, 0.05 and 0.1.

Sample size	Nom.size	LR	$Score$	LM	P_a	P_b	P_{alt}
$n = 100$	0.01	0.0186	0.0187	0.0265	0.0228	0.0561	0.0032
	0.05	0.0764	0.0767	0.0893	0.0822	0.1065	0.0259
	0.1	0.1394	0.1396	0.1507	0.1454	0.1541	0.0655
Sample size	Nom.size	LR	$Score$	LM	P_a	P_b	P_{alt}
$n = 250$	0.01	0.0252	0.0252	0.0277	0.0262	0.0484	0.008
	0.05	0.0784	0.0784	0.0819	0.0802	0.1024	0.0423
	0.1	0.1310	0.1310	0.1343	0.1327	0.1530	0.0857
Sample size	Nom.size	LR	$Score$	LM	P_a	P_b	P_{alt}
$n = 500$	0.01	0.0262	0.0262	0.0273	0.0268	0.0474	0.0133
	0.05	0.0740	0.0740	0.0762	0.0749	0.1000	0.0506
	0.1	0.1281	0.1281	0.1302	0.1291	0.1517	0.0939
Sample size	Nom.size	LR	$Score$	LM	P_a	P_b	P_{alt}
$n = 1000$	0.01	0.0215	0.0215	0.0222	0.0217	0.0403	0.0137
	0.05	0.0663	0.0663	0.0674	0.0669	0.0905	0.0475
	0.1	0.1210	0.1210	0.1216	0.1214	0.1447	0.0895

Table 5.1. Estimated size of the test statistics under correct model specification.

The results present further evidence that commonly-used overidentifying moment restrictions tests are significantly over-sized in finite samples. The results of these tests are similar to RS and Imbens et al. [17] which also considered this model for their simulation studies. Out of the

traditional GEL tests proposed in Smith [33], LR and Score have almost identical performance, with LR delivering slightly better performance for $n = 100$.

The P_a test is competitive with LR and Score, and is less over-sized than LM in all cases. P_b is the worst performing test with estimated size as high as 0.0905 for nominal size 0.05 even at the high sample size of $n = 1000$. Interestingly, for the same model, the simulation results from RS show P_b was less over-sized than P_a when the GEL method employed was empirical likelihood, whereas under exponential tilting estimation, P_b was more over-sized relative to P_a. Overall, this suggests the behaviour of implied probabilities that may lead to differences in performance between P_a and P_b is heavily dependent on the GEL method used.

As in RS, the partition-based goodness-of-fit test P_{alt} is the best performing test with its estimated size being the closest to nominal size under all significance levels and sample sizes considered. There is, however, an under-rejection problem in general, especially for nominated significance levels 0.05 and 0.1. Overall, adapting the P_{alt} statistic to involve data-driven partitions has not harmed the superior size properties suggested by the results in RS.

5.1.4 Power analysis

The primary goal of our analysis under model misspecification is to reveal whether we can successfully estimate the UR, and whether examining differences in goodness-of-fit over regions dictated by the UR improves the power to detect subtle forms of neglected heterogeneity.

To generate misspecification, it is now assumed that for a small portion of the sample, z_1 and z_2 are not generated from a $N(0, 0.16)$ distribution. In particular, those observations for which z_1 and z_2 are both greater than 0.3367, (the 80% quantile of the $N(0, 0.16)$ distribution) are transformed to $z_1 + \kappa$ and $z_2 + \kappa$, respectively. The UR is therefore $\tilde{A} = \{(z_1, z_2) \in \mathbb{R}^2 | z_1 > 0.3367 \text{ and } z_2 > 0.3367\}$, the blue square from Figure 5.1.

Since the transformed variables are not generated from a $N(0, 0.16)$ distribution, the moment conditions no longer hold for $\kappa \neq 0$.

Figure 5.2 shows the rate at which the estimated UR $\hat{A}$ equals $\tilde{A}$ as the level of misspecification κ increases, for varying sample sizes, over 2000 simulations for each experiment. As in Section 4.3 the ability to successfuly pick out the true UR $\tilde{A}$ depends heavily on the sample size and level of misspecification.

Even for the largest level of misspecification and sample size we consider, the UR is only picked out in 72.1% of the simulations. Surprisingly, for low levels of misspecification, higher sample sizes appear to reduce the rate at which the UR is identified.

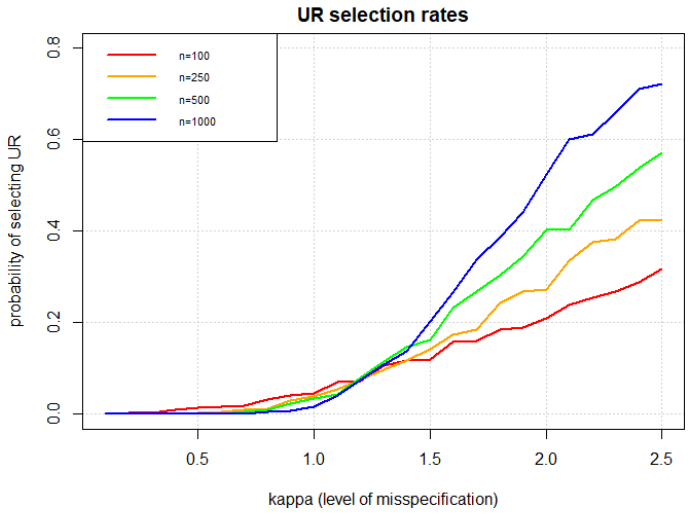
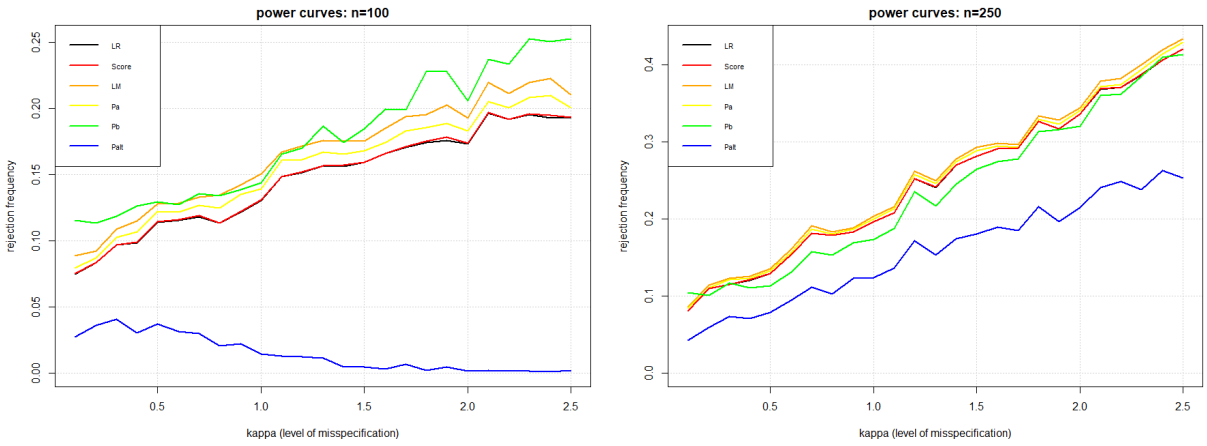


Figure 5.2. UR selection rate for $n = 100, 250, 500, 1000$.

Within these same simulations, we calculated rejection frequencies of the test statistics at a nominated significance level of 0.05. Figure 5.3 presents power results when comparing the test statistics to critical values obtained from the chi-squared distribution $\chi^2_{0.95,1}$. However, since the partition-based goodness-of-fit test was under-sized and the other test statistics were over-sized, we also present size-adjusted power results in Figure 5.4. The size-adjusted rejection rates are calculated by comparing the test statistic against a critical value corresponding to $\chi^2_{0.95-(0.05-\hat{\alpha}),1}$, where $\hat{\alpha}$ are the estimated sizes at the 0.05 level from Table 5.1. For the case of P_b, since for sample sizes $n = 100$, 250 and 500 the estimated size is above 0.1, size-adjusted power would be 0. We cap $\hat{\alpha}$ cap at 0.0999 which therefore presents a distorted and favourable size-adjusted power result for P_b. The other test statistics are unaffected.



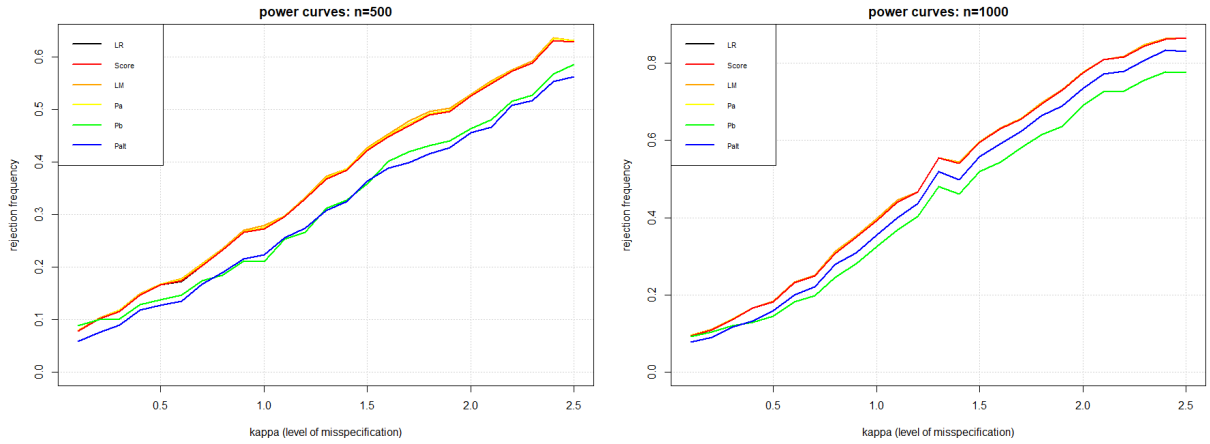
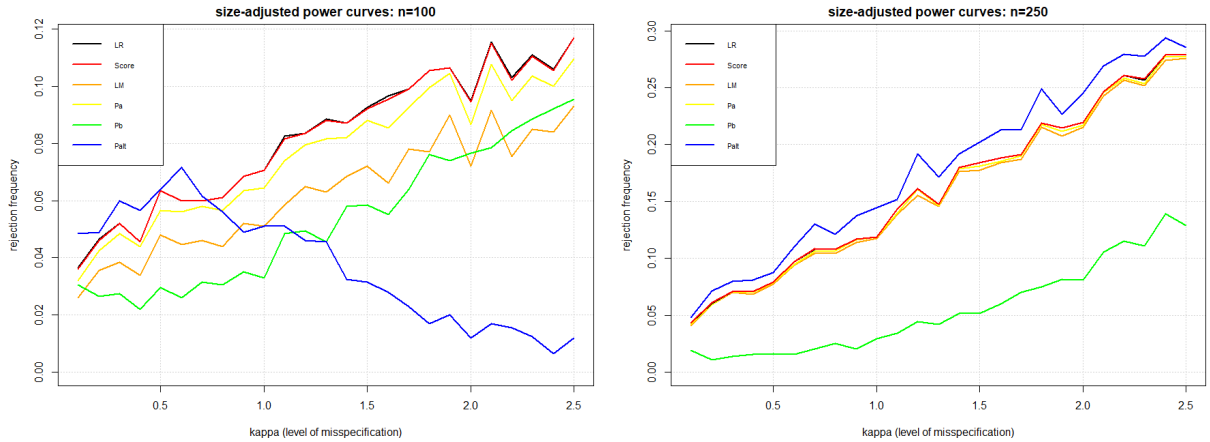


Figure 5.3. Power results for $n = 100, 250, 500, 1000$.

Figure 5.3 shows that for crude power comparisons, P_{alt} needs a large sample size to be competitive with other test statistics. For $n = 100$, P_{alt} has almost zero power to detect model misspecification even when the level of misspecification is high; Figure 5.4 shows that even after accounting for its under-rejection problem, the rejection rate is lower than 0.1 even for a highly misspecified model.

For $n = 250$, Figures 5.3 and 5.4 show that the relatively poorer power results of P_{alt} are down to its under-rejection problem; with its size-adjusted power being greater than other statistics. Likewise, we can also interpret these results as suggesting that any superior power results of other test statistics are down to their inflated type I error rates.

Even though P_b had the most inflated rejection rates under the null, its power results for sample sizes higher than $n = 100$ are relatively poor. Figure 5.4 shows the size-corrected power of P_b is generally distastrous. However, the power results of P_b in RS were less alarming, suggesting again that the performance of the P_b statistic, which is heavily linked to the behaviour of implied probabilities $\{\hat{\pi}_i\}_{i=1}^n$ is largely dependent on the choice of GEL method. For example, empirical likelihood-based P_b statistics may perform well.



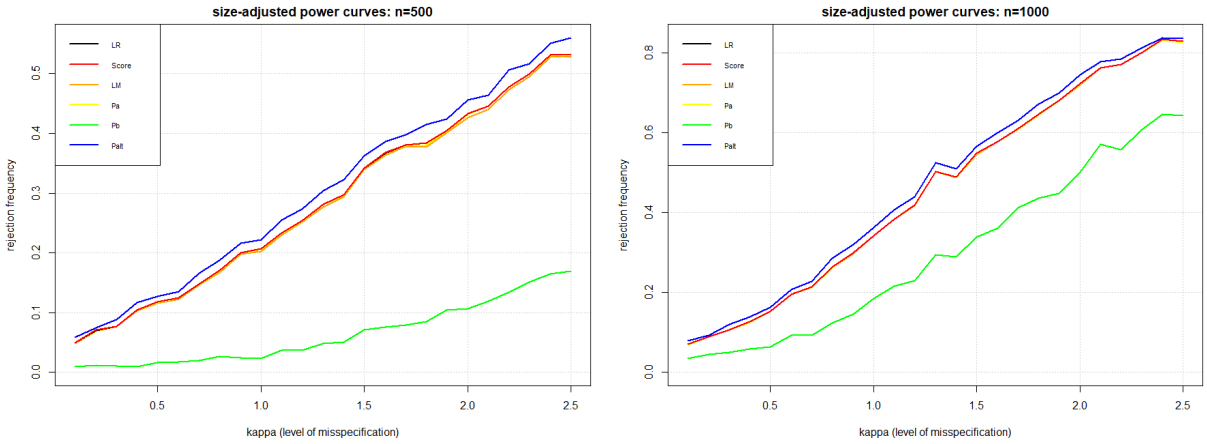


Figure 5.4. Size-adjusted power results for $n = 100, 250, 500, 1000$.

In general, Figure 5.2 with Figures 5.3 and 5.4 suggest low rates of UR selection are linked with relatively low power for detecting neglected heterogeneity, and high rates of UR selection is linked with higher power. However, the sample size appears to be a major factor for both UR selection and good power, and therefore it is difficult to conclude that the proposed data-driven goodness-of-fit test with UR selection improves power. Nevertheless, since for high enough sample sizes the probability of picking out the UR is high, calculation of the UR may be a useful supplementary tool to shed light on a potential neglected heterogeneity problem.

6 Conclusion

This paper derives concentration inequalities for moment condition models that may be used for inference. In particular, VC inequalities are derived which bound the distance between the GEL and empirical distribution functions over partitions of the sample space. These inequalities are non-asymptotic, although the bounds are likely to be conservative. Due to the promising finite-sample performance of the partition-based goodness-of-fit test in RS, the bounds derived here are applied to construct goodness-of-fit tests with data-driven partitions.

The simulation results suggest that while UR estimation may be useful to detect parameter instability, it is unclear if comparing goodness-of-fit over regions determined by the UR will significantly improve power to detect model misspecification. As a result, an interesting development could involve combining usual GMM or GEL test statistics of overidentifying restrictions with a power enhancement component as in Fan et al. [8]. The concentration inequalities derived here can be used to satisfy the 'no distortion' property required for the power enhancement component. However, since existing tests of overidentifying restrictions are typically oversized in finite samples (see the special issue of *Journal of Business and Economic Statistics*, 1996), power enhancement components may further harm finite-sample size properties.

References

[1] Martin Anthony. Aspects of discrete mathematics and probability in the theory of machine learning. *Discrete Applied Mathematics*, 156(6):883–902, 2008.

[2] Martin Anthony and Peter L Bartlett. *Neural Network Learning: Theoretical Foundations*. 1999.

[3] Bertille Antoine, Hélène Bonnal, and Eric Renault. On the efficient use of the informational content of estimating equations: Implied probabilities and Euclidean empirical likelihood. *Journal of Econometrics*, 138(2):461–487, 2007.

[4] P L Bartlett and Gábor Lugosi. An inequality for uniform deviations of sample averages from their means. *Statistics and Probability Letters*, 44(1):55–62, 1999.

[5] Neil M. Davies, Stephanie von Hinke Kessler Scholder, Helmut Farbmacher, Stephen Burgess, Frank Windmeijer, and George Davey Smith. The many weak instruments problem and mendelian randomization. *Statistics in Medicine*, 34(3):454–468, 2015.

[6] Luc Devroye, László Györfi, and Gabor Lugosi. *A Probabilistic Theory of Pattern Recognition*. 1996.

[7] R.M. Dudley. Central Limit Theorems for Empirical Measures. *The Annals of Probability*, 6(6):899–929, 1978.

[8] Jianqing Fan, Yuan Liao, and Jiawei Yao. Power Enhancement in High Dimensional Cross-Sectional Tests. *Econometrica*, 83(4):1497–1541, 2015.

[9] Allan W. Gregory, Jean François Lamarche, and Gregor W. Smith. Information-theoretic estimation of preference parameters: Macroeconomic applications and simulation evidence. *Journal of Econometrics*, 107(1-2):213–233, 2002.

[10] Alain Guay and Jean-François Lamarche. Structural Change Tests Based on Implied Probabilities for Gel Criteria. *Econometric Theory*, 28(06):1186–1228, 2012.

[11] Jinyong Hahn. On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects. *Econometrica*, 66(2):315–331, 1998.

[12] Jinyong Hahn, Whitney K. Newey, and Richard J. Smith. Neglected heterogeneity in moment condition models. *Journal of Econometrics*, 178:86–100, 2014.

[13] Alastair R. Hall. Econometricians Have Their Moments: GMM at 32. *Economic Record*, 91(S1):1–24, 2015.

[14] Lars Peter Hansen. Large sample properties of generalised method of moments estimators. *Econometrica*, 50(4):1029–1054, 1982.

[15] Lars Peter Hansen, John Heaton, and Amir Yaron. Finite-sample properties of some alternative GMM estimators. *Journal of Business & Economic Statistics*, 14(3):262–280, 1996.

[16] Judith K. Hellerstein and Guido W. Imbens. Imposing Moment Restrictions from Auxiliary Data by Weighting. *The Review of Economics and Statistics*, 81(1):1–14, 1999.

[17] Guido W. Imbens, Richard H. Spady, and Phillip Johnson. Information theoretic approaches to inference in moment condition models. *Econometrica*, 66(2):333–357, 1998.

[18] Tadeusz Inglot and Wilbert C M Kallenberg. Moderate deviations of minimum contrast estimators under contamination. *Annals of Statistics*, 31(3):852–879, 2003.

[19] Toru Kitagawa and Aleksey Tetenov. Who should be treated? Empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616, 2018.

[20] Yuichi Kitamura. Asymptotic Optimality of Empirical Likelihood for Testing Moment Restrictions. *Econometrica*, 69(6):1661–1672, 2001.

[21] Yuichi Kitamura, Andres Santos, and Azeem M. Shaikh. On the Asymptotic Optimality of Empirical Likelihood for Testing Moment Restrictions. *Econometrica*, 80(1):413–423, 2012.

[22] Michael R Kosorok. Introduction to Empirical Processes and Semiparametric Inference. *Springer*, pages 1–491, 2008.

[23] Whitney K. Newey. The Asymptotic Variance of Semiparametric Estimators. *Econometrica*, 62(6):1349–1382, 1994.

[24] Whitney K. Newey, Joaquim J S Ramalho, and Richard J. Smith. Identification and Inference for Econometric Models Asymptotic Bias for GMM and GEL Estimators with Estimated Nuisance Parameters. In *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, pages 245–281. 2005.

[25] Whitney K. Newey and Richard J. Smith. Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica*, 72(1):219–255, 2004.

[26] Taisuke Otsu. Large deviations of generalized method of moments and empirical likelihood estimators. *The Econometrics Journal*, 14(2):321–329, 2011.

[27] Taisuke Otsu, Myung Hwan Seo, and Yoon-Jae Whang. Testing for non-nested conditional moment restrictions using unconditional empirical likelihood. *Journal of Econometrics*, 167(2):370–382, 2012.

[28] Taisuke Otsu and Yoon-Jae Whang. Testing for Nonnested Conditional Moment Restrictions Via Conditional Empirical Likelihood. *Econometric Theory*, 27(01):114–153, 2011.

[29] David Pollard. Chaining. In *Update to empirical process book (http://www.stat.yale.edu/~pollard/Books/Mini/Chaining.pdf)*, number November. 2015.

[30] Joaquim J S Ramalho and Richard J. Smith. Goodness of Fit Tests for Moment Condition Models. 2006.

[31] Walter Rudin. *Real and Complex Analysis*. McGraw-Hill Education, third edition, 1987.

[32] Susanne M. Schennach. Point estimation with exponentially tilted empirical likelihood. *Annals of Statistics*, 35(2):634–672, 2007.

[33] Richard J. Smith. Alternative Semi-Parametric Likelihood Approaches to Generalised Method of Moments Estimation. *The Economic Journal*, 107(441):503–519, 1997.

[34] Richard J. Smith. Gel Criteria for Moment Condition Models. *Econometric Theory*, 27(06):1192–1235, 2011.

[35] Aad Van der Vaart and Jon A Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics, 1996.

[36] V Vapnik and Alexey Ya Chervonenkis. On the uniform convergence of relative frequencie of events to their probabilities. *Theory Probab. Appl.*, 16(2):264–280, 1971.

Appendix

A Proof of Theorems 3.1, 3.2, 3.3 and 4.1 and Corollary 5.1

The following notation is used: $g_i(\theta) = g(z_i, \theta)$, $g_n(\theta) = \sum_{i=1}^n g(z_i, \theta)/n$, $g_i^{(k)}(\theta) = g^{(k)}(z_i, \theta)$, $g_n^{(k)}(\theta) = \sum_{i=1}^n g^{(k)}(z_i, \theta)/n$, $G_i(\theta) = G(z_i, \theta)$, $G_n(\theta) = \sum_{i=1}^n G_i(\theta)/n$, $\Omega_i(\theta) = \Omega(z_i, \theta)$, $\Omega_n(\theta) = \sum_{i=1}^n \Omega_i(\theta)/n$, $\Omega(\theta) = \mathbb{E}[g(z, \theta)g(z, \theta)']$.

Proof of Theorem 3.1

Theorem 3.1 states that there exist constants C_{gel} and c_{gel} such that for any $\epsilon > 0$,

$$\mathbb{P}_n\Big(\sup_{A \in \mathcal{A}} \big|\tilde{F}_n(A) - F_n(A)\big| > \epsilon \Big) \leq C_{gel} \exp\{-c_{gel} n\}.$$

Since $\tilde{F}_n(A) - F_n(A) = f_n(\hat{\theta}, \hat{\lambda}(\hat{\theta}); A) = \hat{\lambda}(\hat{\theta})'(\sum_{i=1}^n g_i(\hat{\theta}) \mathbb{I}\{z_i \in A\}/n)$, for any $A \in \mathcal{A}$,

$$\begin{aligned} f_n(\hat{\theta}, \hat{\lambda}(\hat{\theta}); A) &= \hat{\lambda}(\hat{\theta})'\Big(\frac{1}{n}\sum_{i=1}^n \big[g_i(\hat{\theta}) - g_i(\theta_0)\big]\mathbb{I}\{z_i \in A\}\Big) \\ &\quad + \hat{\lambda}(\hat{\theta})'\Big(\frac{1}{n}\sum_{i=1}^n g_i(\theta_0)\mathbb{I}\{z_i \in A\} - \mathbb{E}[g(z, \theta_0)\mathbb{I}\{z \in A\}]\Big) \\ &\quad + \big(\hat{\lambda}(\hat{\theta}) - \lambda(\hat{\theta})\big)'\mathbb{E}[g(z, \theta_0)\mathbb{I}\{z \in A\}] \\ &\quad + \big(\lambda(\hat{\theta}) - 0\big)'\mathbb{E}[g(z, \theta_0)\mathbb{I}\{z \in A\}]. \end{aligned}$$

Using Lemma B.2,

$$\begin{aligned} \mathbb{P}_n\Big(\sup_{A \in \mathcal{A}} \big|f_n(\hat{\theta}, \hat{\lambda}(\hat{\theta}); A)\big| > \epsilon \Big) &\leq \mathbb{P}_n\Big(\sup_{A \in \mathcal{A}} \Big|\hat{\lambda}(\hat{\theta})'\Big(\frac{1}{n}\sum_{i=1}^n \big[g_i(\hat{\theta}) - g_i(\theta_0)\big]\mathbb{I}\{z_i \in A\}\Big)\Big| > \frac{\epsilon}{4}\Big) \\ &\quad + \mathbb{P}_n\Big(\sup_{A \in \mathcal{A}} \Big|\hat{\lambda}(\hat{\theta})'\Big(\frac{1}{n}\sum_{i=1}^n g_i(\theta_0)\mathbb{I}\{z_i \in A\} - \mathbb{E}[g(z, \theta_0)\mathbb{I}\{z \in A\}]\Big)\Big| > \frac{\epsilon}{4}\Big) \\ &\quad + \mathbb{P}_n\Big(\sup_{A \in \mathcal{A}} \Big|\big(\hat{\lambda}(\hat{\theta}) - \hat{\lambda}(\theta_0)\big)'\mathbb{E}[g(z, \theta_0)\mathbb{I}\{z \in A\}]\Big| > \frac{\epsilon}{4}\Big) \\ &\quad + \mathbb{P}_n\Big(\sup_{A \in \mathcal{A}} \Big|\big(\hat{\lambda}(\theta_0) - 0\big)'\mathbb{E}[g(z, \theta_0)\mathbb{I}\{z \in A\}]\Big| > \frac{\epsilon}{4}\Big) \\ &:= J1 + J2 + J3 + J4. \end{aligned} \tag{A.1}$$

Bound each term on the RHS individually. $J1$ is bounded above by

$$\mathbb{P}_n\Big(||\hat{\lambda}(\hat{\theta})|| \times \sup_{A \in \mathcal{A}} \Big|\Big|\frac{1}{n}\sum_{i=1}^n \big[g_i(\hat{\theta}) - g_i(\theta_0)\big]\mathbb{I}\{z_i \in A\}\Big|\Big| > \frac{\epsilon}{4}\Big).$$

By a Taylor expansion, for any $A \in \mathcal{A}$,

$$\Big|\Big|\frac{1}{n}\sum_{i=1}^{n}\big[g_i(\hat{\theta}) - g_i(\theta_0)\big]\mathbb{I}\{z_i \in A\}\Big|\Big| \leq \frac{1}{n}\sum_{i=1}^{n}\Big|\Big|\frac{\partial g_i(\bar{\theta})}{\partial \theta}\Big|\Big| \times ||\hat{\theta} - \theta_0||.$$

By Assumption 2.2(ii) and Lemma D.1(v),

$$\begin{aligned} J1 &\leq \mathbb{P}_n\Big(||\hat{\theta} - \theta_0|| > \frac{\epsilon}{4d_g C_G C_{\lambda,n}}\Big) \\ &\leq C_\theta^{(1)} \exp\{-c_\theta^{(1)} n\} \end{aligned} \tag{A.2}$$

for some $C_\theta^{(1)}, c_\theta^{(1)}$ by Lemma B.5.

By Assumption 2.2(ii), $J2$ is bounded above by

$$\mathbb{P}_n\Big(\sup_{A\in\mathcal{A}}\Big|\Big|\frac{1}{n}\sum_{i=1}^{n} g_i(\theta_0)\mathbb{I}\{z_i \in A\} - \mathbb{E}[g(z,\theta_0)\mathbb{I}\{z \in A\}]\Big|\Big| > \frac{\epsilon}{4C_{\lambda,n}}\Big).$$

For any $A \in \mathcal{A}$,

$$\Big|\Big|\frac{1}{n}\sum_{i=1}^{n} g_i(\theta_0)\mathbb{I}\{z_i \in A\} - \mathbb{E}[g(z,\theta_0)\mathbb{I}\{z \in A\}]\Big|\Big|_1 \leq \sum_{k=1}^{d_g}\Big|\frac{1}{n}\sum_{i=1}^{n} g_i^{(k)}(\theta_0)\mathbb{I}\{z_i \in A\} - \mathbb{E}[g^{(k)}(z,\theta_0)\mathbb{I}\{z \in A\}]\Big|.$$

Using Lemma B.2,

$$J2 \leq \sum_{k=1}^{d_g}\mathbb{P}_n\Big(\sup_{A\in\mathcal{A}}\Big|\frac{1}{n}\sum_{i=1}^{n} g_i^{(k)}(\theta_0)\mathbb{I}\{z_i \in A\} - \mathbb{E}[g^{(k)}(z,\theta_0)\mathbb{I}\{z \in A\}]\Big| > \frac{\epsilon}{4d_g C_{\lambda,n}}\Big).$$

For any general $k \in \{1, ..., d_g\}$, by Lemma C.5,

$$\mathbb{P}_n\Big(\sup_{A\in\mathcal{A}}\Big|\frac{1}{n}\sum_{i=1}^{n} g_i^{(k)}(\theta_0)\mathbb{I}\{z_i \in A\} - \mathbb{E}[g^{(k)}(z,\theta_0)\mathbb{I}\{z \in A\}]\Big| > \epsilon_{1,\star}\Big) \leq 8\Big(\frac{16enC_g}{(2v+1)\epsilon_{1,\star}}\Big)^{2v+1}\exp\Big(-\frac{\epsilon_{1,\star}^2 n}{128C_g^2}\Big)$$

where $\epsilon_{1,\star} = (\epsilon/4d_g C_{\lambda,n})$. Therefore,

$$J2 \leq 8d_g\Big(\frac{16enC_g}{(2v+1)\epsilon_{1,\star}}\Big)^{2v+1}\exp\Big(-\frac{\epsilon_{1,\star}^2 n}{128C_g^2}\Big). \tag{A.3}$$

For $J3$, since under our assumptions, $\hat{\lambda}(\theta)$ is continuously differentiable by the implicit function theorem (see Smith [34], p. 1235),

$$\begin{aligned} \mathbb{P}_n\Big(\sup_{A\in\mathcal{A}}\Big|\big(\hat{\lambda}(\hat{\theta}) - \hat{\lambda}(\theta_0)\big)'\mathbb{E}[g(z,\theta_0)\mathbb{I}\{z \in A\}]\Big| > \frac{\epsilon}{4}\Big) &\leq \mathbb{P}_n\Big(\big|\big|\hat{\lambda}(\hat{\theta}) - \hat{\lambda}(\theta_0)\big|\big| > \frac{\epsilon}{4C_{B,\mathcal{A}}}\Big) \\ &\leq \mathbb{P}_n\Big(\Big|\Big|\frac{\partial\hat{\lambda}(\bar{\theta})}{\partial\theta'}\Big|\Big| \times ||\hat{\theta} - \theta_0|| > \frac{\epsilon}{4C_{B,\mathcal{A}}}\Big) \end{aligned} \tag{A.4}$$

by Assumption 2.2 (viii), and for some $\bar{\theta}$ on a line joining $\hat{\theta}$ and θ_0.

By Smith [34], p. 1235,

$$\frac{\partial \hat{\lambda}(\theta)}{\partial \theta'} = -\Big(\frac{\partial \hat{P}_n(\hat{\lambda}(\theta),\theta)}{\partial \lambda \partial \lambda'}\Big)^{-1} \frac{\partial \hat{P}_n(\hat{\lambda}(\theta),\theta)}{\partial \lambda \partial \theta'} \tag{A.5}$$

where

$$\frac{\partial \hat{P}_n(\hat{\lambda}(\theta),\theta)}{\partial \lambda \partial \lambda'} = \frac{1}{n}\sum_{i=1}^{n} \rho_2(\hat{\lambda}(\theta)' g_i(\theta)) g_i(\theta) g_i(\theta)' \tag{A.6}$$

and

$$\frac{\partial \hat{P}_n(\hat{\lambda}(\theta),\theta)}{\partial \lambda \partial \theta'} = \frac{1}{n}\sum_{i=1}^{n} \rho_2(\hat{\lambda}(\theta)' g_i(\theta)) g_i(\theta)(\hat{\lambda}(\theta)' G_i(\theta)) + \frac{1}{n}\sum_{i=1}^{n} \rho_1(\hat{\lambda}(\theta)' g_i(\theta)) G_i(\theta). \tag{A.7}$$

By Assumption 2.2(vi), due to consistency of $\hat{\theta}$ and hence $\bar{\theta}$, for large enough n,

$$\frac{1}{n}\sum_{i=1}^{n} \rho_2(\hat{\lambda}(\bar{\theta})' g_i(\bar{\theta})) g_i(\bar{\theta}) g_i(\bar{\theta})' \geq \frac{1}{n}\sum_{i=1}^{n} \rho_i^{\Omega}$$

and $\sum_{i=1}^{n} \rho_i^{\Omega}/n$ has minimum eigenvalue δ_ρ. This implies that $\big(\sum_{i=1}^{n} \rho_i^{\Omega}/n\big)^{-1}$ has maximum eigenvalue δ_ρ^{-1}, so that

$$\Big|\Big|\Big(\frac{1}{n}\sum_{i=1}^{n} \rho_2(\hat{\lambda}(\bar{\theta})' g_i(\bar{\theta})) g_i(\bar{\theta}) g_i(\bar{\theta})'\Big)^{-1}\Big|\Big| \leq \Big|\Big|\big(\frac{1}{n}\sum_{i=1}^{n} \rho_i^{\Omega}\big)^{-1}\Big|\Big| \leq \frac{1}{\delta_\rho}. \tag{A.8}$$

By Assumptions 2.2(ii), (vii), and Lemma D.1(i) and (v),

$$\begin{aligned}\Big|\Big|\frac{1}{n}\sum_{i=1}^{n} \rho_2(\hat{\lambda}(\bar{\theta})' g_i(\bar{\theta})) g_i(\bar{\theta})(\hat{\lambda}(\bar{\theta})' G_i(\bar{\theta}))\Big|\Big| &\leq \max_i \Big(|\rho_2(\hat{\lambda}(\bar{\theta})' g_i(\bar{\theta}))|\Big) \times \frac{1}{n}\sum_{i=1}^{n} ||g_i(\bar{\theta})|| \times C_{\lambda,n} \times ||G_i(\bar{\theta})|| \\ &\leq C_{\rho_2} d_g^2 C_{\lambda,n} C_g C_G. \end{aligned} \tag{A.9}$$

Similarly,

$$\Big|\Big|\frac{1}{n}\sum_{i=1}^{n} \rho_1(\hat{\lambda}(\bar{\theta})' g_i(\bar{\theta})) G_i(\bar{\theta})\Big|\Big| \leq C_{\rho_1} d_g C_G. \tag{A.10}$$

Then by $(A.5)$-$(A.10)$,

$$\Big|\Big|\frac{\partial \hat{\lambda}(\bar{\theta})}{\partial \theta'}\Big|\Big| \leq \frac{d_g C_G}{\delta_\rho}\Big(C_{\rho_2} d_g C_{\lambda,n} C_g + C_{\rho_1}\Big) := C_4. \tag{A.11}$$

Substituting into $(A.4)$,

$$J3 \leq \mathbb{P}_n\Big(||\hat{\theta} - \theta_0|| > \frac{\epsilon}{4 C_4 C_{B,\mathcal{A}}}\Big) \leq C_\theta^{(2)} \exp\{-c_\theta^{(2)} n\} \tag{A.12}$$

for some $C_\theta^{(2)}, c_\theta^{(2)}$ by Lemma B.5.

For $J4$, noting that $\lambda(\theta_0) = 0$, by Assumption 2.2(viii),

$$\begin{aligned} J4 &\leq \mathbb{P}_n\Big(\big|\big|\hat{\lambda}(\theta_0) - \lambda_0(\theta_0)\big|\big| > \frac{\epsilon}{4C_{B,\mathcal{A}}}\Big) \\ &\leq C_\lambda^{(1)} \exp\{-c_\lambda^{(1)} n\} \end{aligned} \tag{A.13}$$

for some $C_\lambda^{(1)}$, $c_\lambda^{(1)}$ by Lemma B.4.

Finally, from $(A.1)$, $(A.2)$, $(A.3)$, $(A.12)$ and $(A.13)$, the above arguments establish that there exist constants C_{gel}, c_{gel} such that Theorem 3.1 is satisfied. □

Proof of Theorem 3.2

Let $Q_n(\theta) = g_n(\theta)'\Omega_n^{-1}(\theta)g_n(\theta)$ and $Q_0(\theta) = \mathbb{E}[g(z,\theta)]'\Omega^{-1}(\theta)\mathbb{E}[g(z,\theta)]$. The CU-GMM estimator solves

$$\hat{\theta} = \arg\min_{\theta\in\Theta} g_n(\theta)'\Omega_n^{-1}(\theta)g_n(\theta).$$

Let $\delta = \inf_{\theta\in\Theta, |\theta-\theta_0|>\epsilon} Q_0(\theta) - Q_0(\theta_0) > 0$. Then arguing as in Lemma A.2 of Otsu [26],

$$\mathbb{P}_n\big(||\hat{\theta} - \theta_0|| > \epsilon\big) \leq \mathbb{P}_n\Big(\sup_{\theta\in\Theta}\big|Q_n(\theta) - Q_0(\theta)\big| > \frac{\delta}{3}\Big). \tag{A.14}$$

Define the following events

$$\begin{aligned} E_{Q,n} &= \Big\{\sup_{\theta\in\Theta}\big|Q_n(\theta) - Q_0(\theta)\big| > \frac{\delta}{3}\Big\} \\ E_{\Omega,n} &= \Big\{\sup_{\theta\in\Theta}||\Omega_n^{-1}(\theta) - \Omega^{-1}(\theta)||_1 \leq 1\Big\}. \end{aligned}$$

Let the event $E_{\Omega,n}$ hold, and define $\mathbb{P}_n^{|}(.) = \mathbb{P}_n(.|E_{\Omega,n})$, that is, the probability of an event conditional on $E_{\Omega,n}$.

Now, by T and CS,

$$\begin{aligned} \big|Q_n(\theta) - Q_0(\theta)\big| \leq\ & ||g_n(\theta) - \mathbb{E}[g(z,\theta)]|| \times ||\Omega_n^{-1}(\theta)|| \times ||g_n(\theta)|| \\ & + ||\mathbb{E}[g(z,\theta)]|| \times ||\Omega_n^{-1}(\theta) - \Omega^{-1}(\theta)|| \times ||g_n(\theta)|| \\ & + ||\mathbb{E}[g(z,\theta)]|| \times ||\Omega^{-1}(\theta)|| \times ||g_n(\theta) - \mathbb{E}[g(z,\theta)]||. \end{aligned}$$

Then using Lemma B.2, by T,

$$\mathbb{P}_n^{|}\Big(\sup_{\theta\in\Theta}\big|Q_n(\theta)-Q_0(\theta)\big|>\frac{\delta}{3}\Big) \leq \mathbb{P}_n^{|}\Big(\sup_{\theta\in\Theta}||g_n(\theta)-\mathbb{E}[g(z,\theta)]||>\frac{\delta\delta_{\Omega,\min}}{9d_gC_g(1+\delta_{\Omega,\min})}\Big) \quad (A.15)$$

$$+\mathbb{P}_n^{|}\Big(\sup_{\theta\in\Theta}||\Omega_n^{-1}(\theta)-\Omega^{-1}(\theta)||>\frac{\delta}{9d_g^2C_g^2}\Big) \quad (A.16)$$

$$+\mathbb{P}_n^{|}\Big(\sup_{\theta\in\Theta}||g_n(\theta)-\mathbb{E}[g(z,\theta)]||>\frac{\delta}{9d_gC_g\delta_{\Omega,\min}}\Big) \quad (A.17)$$

where the first term on the RHS follows from Lemmata D.1(i) and D.2(ii), and the event $E_{\Omega,n}$ holding. The second term on the RHS follows from Assumption 2.1(ii), Lemmata D.1(i) and D.1(ii), and the third term on the RHS follows from Lemmata D.1(ii) and D.2(i).

For $(A.15)$, note that $||g_n(\theta)-\mathbb{E}[g(z,\theta)]||_1=\sum_{k=1}^{d_g}|g_n^{(k)}(\theta)-\mathbb{E}[g^{(k)}(z,\theta)]|$. Therefore, for any $\kappa>0$, using Lemma B.2,

$$\begin{aligned}\mathbb{P}_n^{|}\Big(\sup_{\theta\in\Theta}||g_n(\theta)-\mathbb{E}[g(z,\theta)]||_1>\kappa\Big) &= \mathbb{P}_n^{|}\Big(\sum_{k=1}^{d_g}\sup_{\theta\in\Theta}|g_n^{(k)}(\theta)-\mathbb{E}[g^{(k)}(z,\theta)]|>\kappa\Big)\\ &\leq \sum_{k=1}^{d_g}\mathbb{P}_n^{|}\Big(\sup_{\theta\in\Theta}|g_n^{(k)}(\theta)-\mathbb{E}[g^{(k)}(z,\theta)]|>\frac{\kappa}{d_g}\Big)\end{aligned}$$

For any $k\in\{1,...,d_g\}$, $\sup_{\theta\in\Theta}\sup_{z\in\mathcal{Z}}|g^{(k)}(z,\theta)|\leq C_g$. Let $\mathcal{F}_{|z}=\{g(z,\theta):\theta\in\Theta\}$ be a class of functions indexed by θ for fixed $z=(z_1,...,z_n)$. Arguing as in Proof of Lemma C.5 [1], for $n>8d_gC_g^2/\kappa$,

$$\mathbb{P}_n^{|}\Big(\sup_{\theta\in\Theta}|g_n^{(k)}(\theta)-\mathbb{E}g^{(k)}(z,\theta)|>\frac{\kappa}{d_g}\Big)\leq 8\mathbb{E}_z(\mathcal{N}(\frac{\kappa}{8d_g},\mathcal{F}_{|z},d_1))\exp\Big\{-\frac{\kappa^2n}{128d_g^2C_g^2}\Big\}. \quad (A.18)$$

For any fixed $\bar\theta\in\Theta$, $\sup_{z\in\mathcal{Z}}|g^{(k)}(z,\theta)-g^{(k)}(z,\bar\theta)|\leq d_\theta C_G||\theta-\bar\theta||=C_Gd_1(\theta,\bar\theta)$, by Assumption 2.1(iii) where d_1 is the d_1-distance $d_1(\theta,\bar\theta)=\sum_{j=1}^{d_\theta}|\theta_j-\bar\theta_j|$. Then C_G is an envelope for the class $\{g^{(j)}(z,\theta)-g^{(j)}(z,\bar\theta):\theta\in\Theta\}$. For such a class the covering numbers are bounded by bracketing numbers, see p.84 of Van der Vaart and Wellner [35]. By applying Theorem 2.7.11 of Van der Vaart and Wellner [35], the bracketing numbers for this class satisfy, for any $\kappa_\star>0$,

$$\mathcal{N}\Big(\kappa_\star,\mathcal{F},d_1\Big)\leq\mathcal{N}_{[]}(\kappa_\star,\mathcal{F},l_1)\leq\mathcal{N}\Big(\frac{\kappa_\star}{2C_G},\Theta,d_1\Big) \quad (A.19)$$

where l_1 is the l_1-norm. Now if Θ is a ball in $\mathbb{R}^{d_\theta}$ with radius R, by Example 9 of Pollard [29], for any $0<\kappa_{\star,2}\leq R$,

$$\mathcal{D}(\kappa_{\star,2},\Theta,d_1)\leq\Big(\frac{3R}{\kappa_{\star,2}}\Big)^{d_\theta}. \quad (A.20)$$

Note that covering numbers are bounded above by the corresponding packing numbers, p. 98

[1] Also see Theorem 29.1 of Devroye et al. [6].

of Van der Vaart and Wellner [35], so that

$$\mathcal{N}\Big(\frac{\kappa_\star}{2C_G}, \Theta, d_1\Big) \leq \mathcal{D}(\frac{\kappa_\star}{2C_G}, \Theta, d_1). \tag{A.21}$$

Using $(A.20)$,

$$\mathcal{D}(\frac{\kappa_\star}{2C_G}, \Theta, d_1) \leq \Big(\frac{6C_G R}{\kappa_\star}\Big)^{d_\theta}. \tag{A.22}$$

Then from $(A.18)$, $(A.19)$, $(A.21)$ and $(A.22)$, setting $\kappa_\star = \kappa/8d_g$,

$$\mathbb{P}_n^{|}\Big(\sup_{\theta\in\Theta}|g_n^{(k)}(\theta) - \mathbb{E}g^{(k)}(z,\theta)| > \frac{\kappa}{d_g}\Big) \leq 8\Big(\frac{48 d_g C_G R}{\kappa}\Big)^{d_\theta} \exp\Big\{-\frac{\kappa^2 n}{128 d_g^2 C_g^2}\Big\}. \tag{A.23}$$

Since this holds for any $k \in \{1, ..., d_g\}$, since from above $\mathbb{P}_n^{|}(\sup_{\theta\in\Theta}||g_n(\theta) - \mathbb{E}[g(z,\theta)]||_1 > \kappa) \leq \sum_{k=1}^{d_g}\mathbb{P}_n^{|}(\sup_{\theta\in\Theta}|g_n^{(k)}(\theta) - \mathbb{E}[g^{(k)}(z,\theta)]| > \kappa/d_g)$,

$$\mathbb{P}_n^{|}\Big(\sup_{\theta\in\Theta}||g_n(\theta) - \mathbb{E}[g(z,\theta)]||_1 > \kappa\Big) \leq 8d_g\Big(\frac{48 d_g C_G R}{\kappa}\Big)^{d_\theta} \exp\Big\{-\frac{\kappa^2 n}{128 d_g^2 C_g^2}\Big\}. \tag{A.24}$$

Note that for an $d_g \times d_g$ matrix X, $||X||_1 = \max_j \sum_{i=1}^{d_g}|x_{ij}|$. Therefore,

$$||\Omega_n(\theta) - \Omega(\theta)||_1 \leq \max_{,j}\sum_{k=1}^{d_g}\Big|\Big(\frac{1}{n}\sum_{i=1}^{n} g^{(k)}(z_i,\theta)g^{(j)}(z_i,\theta)\Big) - \mathbb{E}[g^{(k)}(z,\theta)g^{(j)}(z,\theta)]\Big|.$$

Therefore, by Lemma B.2,

$$\begin{aligned}
\mathbb{P}_n^{|}(\sup_{\theta\in\Theta}||\Omega_n(\theta) - \Omega(\theta)||_1 > \kappa) &\leq \sum_{k=1}^{d_g}\mathbb{P}_n^{|}\Big(\max_j\sup_{\theta\in\Theta}\Big|\frac{1}{n}\sum_{i=1}^{n} g^{(k)}(z_i,\theta)g^{(j)}(z_i,\theta) \\
&\quad -\mathbb{E}[g^{(k)}(z,\theta)g^{(j)}(z,\theta)]\Big| > \frac{\kappa}{d_g}\Big) \\
&\leq d_g^2\max_{j,k}\mathbb{P}_n^{|}\Big(\sup_{\theta\in\Theta}\Big|\frac{1}{n}\sum_{i=1}^{n} g^{(k)}(z_i,\theta)g^{(j)}(z_i,\theta) \\
&\quad -\mathbb{E}[g^{(k)}(z,\theta)g^{(j)}(z,\theta)]\Big| > \frac{\kappa}{d_g}\Big)
\end{aligned} \tag{A.25}$$

where the second inequality follows from the union bound.

For any $j, k \in \{1, ..., d_g\}$, note that $\sup_{\theta\in\Theta}|g^{(k)}(z,\theta)g^{(j)}(z,\theta)| \leq C_g^2$ by Assumption 2.1(ii). Also, for any $j, k \in \{1, ..., d_g\}$, for any fixed $\bar{\theta} \in \Theta$, by Assumptions 2.1(ii), (iii),

$$\sup_{z\in\mathcal{Z}}|g^{(k)}(z,\theta)g^{(j)}(z,\theta) - g^{(k)}(z,\bar{\theta})g^{(j)}(z,\bar{\theta})| \leq 2C_G C_g||\theta - \bar{\theta}||.$$

Then using the same covering arguments used to derive $(A.23)$,

$$\mathbb{P}_n^{|}\Big(\sup_{\theta\in\Theta}\Big|\frac{1}{n}\sum_{i=1}^{n} g^{(j)}(z_i,\theta)g^{(k)}(z_i,\theta) - \mathbb{E}[g^{(j)}(z,\theta)g^{(k)}(z,\theta)]\Big| > \frac{\kappa}{d_g}\Big) \\ \leq 8\Big(\frac{96 d_g C_g C_G R}{\kappa}\Big)^{d_\theta}\exp\Big\{-\frac{\kappa^2 n}{128 d_g^2 C_g^4}\Big\}.$$

Since this holds for any $j,k\in\{1,...,d_g\}$, by the second inequality of $(A.25)$,

$$\mathbb{P}_n^{|}(\sup_{\theta\in\Theta}||\Omega_n(\theta)-\Omega(\theta)||_1 > \kappa) \leq 8d_g^2\Big(\frac{96 d_g C_g C_G R}{\kappa}\Big)^{d_\theta}\exp\Big\{-\frac{\kappa^2 n}{128 d_g^2 C_g^4}\Big\}. \tag{A.26}$$

Then by Assumption 2.3, for any $\kappa_3>0$, and using $(A.26)$ with $\kappa=\kappa_3/c_\Omega^{(1)}$,

$$\mathbb{P}_n^{|}(\sup_{\theta\in\Theta}||\Omega_n^{-1}(\theta)-\Omega^{-1}(\theta)|| > \kappa_3) \leq 8C_\Omega^{(1)}d_g^2\Big(\frac{96 d_g c_\Omega^{(1)} C_g C_G R}{\kappa_3}\Big)^{d_\theta}\exp\Big\{-\frac{\kappa_3^2 n}{128 (c_\Omega^{(1)})^2 d_g^2 C_g^4}\Big\}. \tag{A.27}$$

Now by $(A.15)-(A.17)$, substituting in bounds $(A.24)$ with $\kappa_1=\delta\delta_{\Omega,\min}/(9d_gC_g(1+\delta_{\Omega,\min}))$ and then $\kappa_2=\delta/(9d_gC_g\delta_{\Omega,\min})$, and substituting bound $(A.27)$ with $\kappa_3=\delta/(9d_g^2C_g^2)$,

$$\begin{aligned}\mathbb{P}_n^{|}\Big(\sup_{\theta\in\Theta}|Q_n(\theta)-Q_0(\theta)| > \frac{\delta}{3}\Big) \;\leq\; & 8d_g\Big(\frac{48 d_g C_G R}{\kappa_1}\Big)^{d_\theta}\exp\Big\{-\frac{\kappa_1^2 n}{128 d_g^2 C_g^2}\Big\} + 8d_g\Big(\frac{48 d_g C_G R}{\kappa_2}\Big)^{d_\theta}\exp\Big\{-\frac{\kappa_2^2 n}{128 d_g^2 C_g^2}\Big\} \\ & + 8C_\Omega^{(1)}d_g^2\Big(\frac{96 d_g c_\Omega^{(1)} C_g C_G R}{\kappa_3}\Big)^{d_\theta}\exp\Big\{-\frac{\kappa_3^2 n}{128 (c_\Omega^{(1)})^2 d_g^2 C_g^4}\Big\}.\end{aligned}$$

To simplify notation further, define the following constants.

$$\omega_1 = 3\max\Big\{8d_g\Big(\frac{48 d_g C_G R}{\kappa_1}\Big)^{d_\theta}, 8d_g\Big(\frac{48 d_g C_G R}{\kappa_2}\Big)^{d_\theta}, 8C_\Omega^{(1)}d_g^2\Big(\frac{96 d_g c_\Omega^{(1)} C_g C_G R}{\kappa_3}\Big)^{d_\theta}\Big\} \tag{A.28}$$

$$\omega_2 = \min\Big\{\frac{\bar\kappa_1^2}{128 d_g^2 C_g^2}, \frac{\bar\kappa_2^2}{128 d_g^2 C_g^2}, \frac{\bar\kappa_3^2}{128 (c_\Omega^{(1)})^2 d_g^2 C_g^4}\Big\} \tag{A.29}$$

where $\bar\kappa_1=\delta_{\Omega,\min}/(9d_gC_g(1+\delta_{\Omega,\min}))$, $\bar\kappa_2=1/(9d_gC_g\delta_{\Omega,\min})$ and $\bar\kappa_3=1/(9d_g^2C_g^2)$.

Since event $\{E_{\Omega,n}\}$ holds,

$$\mathbb{P}_n\big(E_{Q,n}|\,E_{\Omega,n}\big) \leq \omega_1\exp\{-\omega_2\delta^2 n\}. \tag{A.30}$$

Using Lemma B.1,

$$\begin{aligned}\mathbb{P}_n(E_{Q,n}) &\leq \mathbb{P}_n(E_{Q,n}\cap E_{\Omega,n}) + \mathbb{P}_n(E_{\Omega,n}^c) \\ &= \mathbb{P}_n(E_{Q,n}|E_{\Omega,n})\mathbb{P}_n(E_{\Omega,n}) + \mathbb{P}_n(E_{\Omega,n}^c) \\ &\leq \mathbb{P}_n(E_{Q,n}|E_{\Omega,n}) + \mathbb{P}_n(E_{\Omega,n}^c).\end{aligned} \tag{A.31}$$

Therefore,

$$\mathbb{P}_n(E_{Q,n}) \leq \omega_1 \exp\{-\omega_2 \delta^2 n\} + \omega_3 \exp\Big\{ -\omega_4 n \Big\} \tag{A.32}$$

where ω_1 and ω_2 are defined in $(A.28)$ and $(A.29)$, respectively. And from $(A.27)$ with $\kappa_3 = 1$,

$$\begin{aligned} \omega_3 &= 8C_\Omega^{(1)} d_g^2 (96 d_g c_\Omega^{(1)} C_g C_G R)^{d_\theta} && \text{(A.33)} \\ \omega_4 &= \frac{1}{128 (c_\Omega^{(1)})^2 d_g^2 C_g^4}. && \text{(A.34)} \end{aligned}$$

The result follows by noting that $\mathbb{P}_n(||\hat{\theta} - \theta_0|| > \epsilon) \leq \mathbb{P}_n(E_{Q,n})$ by $(A.14)$. □

Proof of Theorem 3.3

Similar arguments to those of the proof of Theorem 3.2 are used. Consider the event

$$E_{\Omega,n} = \Big\{ \sup_{\theta \in \Theta} ||\Omega_n^{-1}(\theta) - \Omega^{-1}(\theta)||_1 \leq 1 \Big\}.$$

Let $\mathbb{P}_n^{|}(.) = \mathbb{P}_n(.|E_{\Omega,n})$. Also let $f_n(\theta, A) = g_n(\theta)' \Omega_n^{-1}(\theta) \big(\frac{1}{n} \sum_{i=1}^n g(z_i, \theta) \mathbb{I}\{z_i \in A\} \big)$. Then, since $\mathbb{E}[g(z, \theta_0)] = 0$,

$$\begin{aligned} f_n(\hat{\theta}, A) &= [g_n(\hat{\theta}) - g_n(\theta_0)]' \Omega_n^{-1}(\hat{\theta}) \Big(\frac{1}{n} \sum_{i=1}^n g(z_i, \hat{\theta}) \mathbb{I}\{z_i \in A\} \Big) \\ &\quad + \big(g_n(\theta_0) - \mathbb{E}[g(z, \theta_0)] \big)' \Omega_n^{-1}(\hat{\theta}) \Big(\frac{1}{n} \sum_{i=1}^n g(z_i, \hat{\theta}) \mathbb{I}\{z_i \in A\} \Big) \end{aligned}$$

or,

$$\begin{aligned} f_n(\hat{\theta}, A) &= g_n(\hat{\theta})' \Omega_n^{-1}(\hat{\theta}) \Big[\frac{1}{n} \sum_{i=1}^n (g_i(\hat{\theta}) - g_i(\theta_0)) \mathbb{I}\{z_i \in A\} \Big] \\ &\quad + g_n(\hat{\theta})' \Omega_n^{-1}(\hat{\theta}) \Big[\frac{1}{n} \sum_{i=1}^n g_i(\theta_0) \mathbb{I}\{z_i \in A\} - \mathbb{E}[g(z, \theta_0) \mathbb{I}\{z \in A\}] \Big] \\ &\quad + \big[g_n(\hat{\theta}) - g_n(\theta_0) \big]' \Omega_n^{-1}(\hat{\theta}) \mathbb{E}[g(z, \theta_0) \mathbb{I}\{z \in A\}] \\ &\quad + \big[g_n(\theta_0) - \mathbb{E}[g(z, \theta_0)] \big]' \Omega_n^{-1}(\hat{\theta}) \mathbb{E}[g(z, \theta_0) \mathbb{I}\{z \in A\}]. \end{aligned}$$

Therefore, by T and CS,

$$\begin{aligned} |f_n(\hat{\theta}, A)| &\leq ||g_n(\hat{\theta}) - g_n(\theta_0)|| \times ||\Omega_n^{-1}(\hat{\theta})|| \times \Big|\Big| \frac{1}{n} \sum_{i=1}^n g(z_i, \hat{\theta}) \mathbb{I}\{z_i \in A\} \Big|\Big| \\ &\quad + ||g_n(\theta_0) - \mathbb{E}[g(z, \theta_0)]|| \times |||\Omega_n^{-1}(\hat{\theta})|| \times \Big|\Big| \frac{1}{n} \sum_{i=1}^n g(z_i, \hat{\theta}) \mathbb{I}\{z_i \in A\} \Big|\Big| \\ &:= T_1^A + T_2^A \end{aligned}$$

or,

$$\begin{aligned}
|f_n(\hat{\theta}, A)| &\leq ||g_n(\hat{\theta})|| \times ||\Omega_n^{-1}(\hat{\theta})|| \times \Big|\Big|\frac{1}{n}\sum_{i=1}^{n}(g_i(\hat{\theta}) - g_i(\theta_0))\mathbb{I}\{z_i \in A\}\Big|\Big| \\
&\quad + ||g_n(\hat{\theta})|| \times ||\Omega_n^{-1}(\hat{\theta})|| \times \Big|\Big|\frac{1}{n}\sum_{i=1}^{n} g_i(\theta_0)\mathbb{I}\{z_i \in A\} - \mathbb{E}[g(z,\theta_0)\mathbb{I}\{z \in A\}]\Big|\Big| \\
&\quad + ||g_n(\hat{\theta}) - g_n(\theta_0)|| \times ||\Omega_n^{-1}(\hat{\theta})|| \times ||\mathbb{E}[g(z,\theta_0)\mathbb{I}\{z \in A\}]|| \\
&\quad + ||g_n(\theta_0) - \mathbb{E}[g(z,\theta_0)]|| \times ||\Omega_n^{-1}(\hat{\theta})|| \times ||\mathbb{E}[g(z,\theta_0)\mathbb{I}\{z \in A\}]|| \\
&:= T_1^B + T_2^B + T_3^B + T_4^B.
\end{aligned}$$

Using the RHS terms of the above inequalities, for any $A \in \mathcal{A}$,

$$\mathbb{P}_n^{|}\big(|f_n(\hat{\theta}, A)| > \epsilon\big) \leq \min\Big\{\mathbb{P}_n^{|}\big(T_1^A + T_2^A > \epsilon\big), \mathbb{P}_n^{|}\big(T_1^B + T_2^B + T_3^B + T_4^B > \epsilon\big)\Big\}. \quad (A.35)$$

First consider probability bounds for $T_1^A + T_2^A$. By a Taylor expansion, for some $\bar{\theta}$ on a line joining $\hat{\theta}$ and θ_0,

$$\begin{aligned}
\Big|\Big|\frac{1}{n}\sum_{i=1}^{n}(g_i(\hat{\theta}) - g_i(\theta_0))\Big|\Big| &\leq \frac{1}{n}\sum_{i=1}^{n}\Big|\Big|\frac{\partial g_i(\bar{\theta})}{\partial \theta}\Big|\Big| \times ||\hat{\theta} - \theta_0|| \\
&\leq d_g C_G ||\hat{\theta} - \theta_0|| \qquad (A.36)
\end{aligned}$$

by Lemma D.1(v).

Also from Lemma D.1(iii), $||\frac{1}{n}\sum_{i=1}^{n} g(z_i, \hat{\theta})\mathbb{I}\{z_i \in A\}|| \leq d_g C_g$. Hence, using Lemmata B.2, D.1(i), (ii), from $(A.36)$ and T_1^A and T_2^A,

$$\begin{aligned}
\mathbb{P}_n^{|}\Big(|f_n(\hat{\theta}, A)| > \epsilon\Big) &\leq \mathbb{P}_n^{|}\Big(||\hat{\theta} - \theta_0|| > \frac{\delta_{\Omega,min}\epsilon}{2(1+\delta_{\Omega,min})d_g^2 C_G C_g}\Big) \qquad (A.37) \\
&\quad + \mathbb{P}_n^{|}\Big(||g_n(\theta_0) - \mathbb{E}[g(z,\theta_0)]|| > \frac{\delta_{\Omega,min}\epsilon}{2(1+\delta_{\Omega,min})d_g C_g}\Big). \qquad (A.38)
\end{aligned}$$

Bound the RHS of $(A.37)$ by the bound from Theorem 3.2,

$$\mathbb{P}_n^{|}\Big(||\hat{\theta} - \theta_0|| > \epsilon^\star\Big) \leq \omega_1 \exp\{-\omega_2 \delta^2(\epsilon^\star)n\} + \omega_3 \exp\Big\{-\omega_4 n\Big\} \qquad (A.39)$$

where $\epsilon^\star = \frac{\delta_{\Omega,\min}\epsilon}{2(1+\delta_{\Omega,\min})d_g^2 C_G C_g}$. Note that $\delta(\epsilon^\star) > 0$ in general depends on $\epsilon^\star$. The constants ω_1, ω_2, ω_3 and ω_4 are defined in $(A.28)$, $(A.29)$, $(A.33)$ and $(A.34)$.

Now, by Lemma B.2, for any $\epsilon^{\star\star} > 0$,

$$\begin{aligned}\mathbb{P}_n^{|}\big(||g_n(\theta_0) - \mathbb{E}[g(z,\theta_0)]||_1 > \epsilon^{\star\star}\Big) &= \mathbb{P}_n^{|}\Big(\sum_{k=1}^{d_g} |g_n^{(k)}(\theta_0) - \mathbb{E}[g^{(k)}(z,\theta_0)]| > \epsilon^{\star\star}\Big) \\ &\leq \sum_{k=1}^{d_g} \mathbb{P}_n^{|}\Big(|g_n^{(k)}(\theta_0) - \mathbb{E}[g^{(k)}(z,\theta_0)]| > \epsilon^{\star\star}/d_g\Big).\end{aligned}$$

For any $k \in \{1, ..., d_g\}$, since $\max_{1\leq i\leq n} |g^{(k)}(z_i,\theta_0)| \leq C_g$, Lemma B.3 implies

$$\mathbb{P}_n^{|}\big(|g_n^{(k)}(\theta_0) - \mathbb{E}[g^{(k)}(z,\theta_0)]| > \epsilon^{\star\star}/d_g\big) \leq 2\exp\Big(\frac{-n(\epsilon^{\star\star})^2}{2d_g^2 C_g^2}\Big).$$

Therefore, by Lemma B.2,

$$\mathbb{P}_n^{|}\big(||g_n(\theta_0) - \mathbb{E}[g(z,\theta_0)]||_1 > \epsilon^{\star\star}\Big) \leq 2d_g\exp\Big(\frac{-n(\epsilon^{\star\star})^2}{2d_g^2 C_g^2}\Big). \tag{A.40}$$

From $(A.37)$, $(A.38)$, $(A.39)$ and $(A.40)$,

$$\begin{aligned}\mathbb{P}_n^{|}\big(T_1^A + T_2^A > \epsilon\big) &\leq \omega_1\exp\{-\omega_2\delta^2(\epsilon^\star)n\} + \omega_3\exp\Big\{-\omega_4 n\Big\} \\ &\quad + 2d_g\exp\Big(\frac{-n(\epsilon^{\star\star})^2}{2d_g^2 C_g^2}\Big)\end{aligned}$$

where $\epsilon^\star = \frac{\delta_{\Omega,\min}\epsilon}{2(1+\delta_{\Omega,\min})d_g^2 C_G C_g}$ and $\epsilon^{\star\star} = \frac{\delta_{\Omega,\min}\epsilon}{2(1+\delta_{\Omega,\min})d_g C_g}$, and constants ω_1, ω_2, ω_3 and ω_4 are defined in $(A.28)$, $(A.29)$, $(A.33)$ and $(A.34)$.

Now, consider the following probability bound for $T_1^B + T_2^B + T_3^B + T_4^B$.

For T_1^B, by T and CS, for some $\bar\theta$ on a line joining $\hat\theta$ and θ_0,

$$\begin{aligned}\Big|\Big|\frac{1}{n}\sum_{i=1}^n (g_i(\hat\theta) - g_i(\theta_0))\mathbb{I}\{z_i \in A\}\Big|\Big| &\leq \frac{1}{n}\sum_{i=1}^n \Big|\Big|\frac{\partial g_i(\bar\theta)}{\partial\theta}\Big|\Big| \times ||\hat\theta - \theta_0|| \\ &\leq d_g C_G||\hat\theta - \theta_0|| \end{aligned}\tag{A.41}$$

using Lemma D.1(v). Now using Lemma D.1(i) and Lemma D.2(ii) and $(A.41)$, the probability that T_1^B exceeds $\epsilon/4$ is

$$\begin{aligned}\mathbb{P}_n^{|}(T_1^B > \epsilon/4) &\leq \mathbb{P}_n^{|}\Big(||\hat\theta - \theta_0|| > \frac{\epsilon\delta_{\Omega,min}}{4d_g^2 C_G C_g(1+\delta_{\Omega,min})}\Big) &\text{(A.42)}\\ &\leq \omega_1\exp\{-\omega_2\delta^2(\epsilon_a)n\} + \omega_3\exp\Big\{-\omega_4 n\Big\} &\text{(A.43)}\end{aligned}$$

where $\epsilon_a = \epsilon\delta_{\Omega,\min}/(4d_g^2 C_G C_g(1+\delta_{\Omega,\min}))$, by Theorem 3.2. The constants ω_1, ω_2, ω_3 and ω_4 are defined in $(A.28)$, $(A.29)$, $(A.33)$ and $(A.34)$.

For T_2^B,

$$\sup_{A\in\mathcal{A}}\Big|\Big|\frac{1}{n}\sum_{i=1}^{n} g_i(\theta_0)\mathbb{I}\{z_i\in A\}-\mathbb{E}[g(z,\theta_0)\mathbb{I}\{z\in A\}]\Big|\Big|_1 = \sum_{k=1}^{d_g}\sup_{A\in\mathcal{A}}\Big|\frac{1}{n}\sum_{i=1}^{n} g_i^{(k)}(\theta_0)\mathbb{I}\{z_i\in A\}-\mathbb{E}[g^{(k)}(z,\theta_0)\mathbb{I}\{z\in A\}]\Big|.$$

Using Lemma B.2, the probability that the RHS exceeds any $\epsilon>0$ is bounded above by

$$\sum_{k=1}^{d_g}\mathbb{P}_n^{|}\Big(\sup_{A\in\mathcal{A}}\Big|\frac{1}{n}\sum_{i=1}^{n} g_i^{(k)}(\theta_0)\mathbb{I}\{z_i\in A\}-\mathbb{E}[g^{(k)}(z,\theta_0)\mathbb{I}\{z\in A\}]\Big|>\frac{\epsilon}{d_g}\Big).$$

Applying Lemma C.1 and Lemma C.5, for any $k\in\{1,...,d_g\}$,

$$\mathbb{P}_n^{|}\Big(\sup_{A\in\mathcal{A}}\Big|\frac{1}{n}\sum_{i=1}^{n} g_i^{(k)}(\theta_0)\mathbb{I}\{z_i\in A\}-\mathbb{E}[g^{(k)}(z,\theta_0)\mathbb{I}\{z\in A\}]\Big|>\frac{\epsilon}{d_g}\Big)\le 8\Big(\frac{16end_gC_g}{(2v+1)\epsilon}\Big)^{2v+1}\exp\Big(-\frac{\epsilon^2 n}{128C_g^2d_g^2}\Big).$$

Therefore,

$$\mathbb{P}_n^{|}\Big(\sup_{A\in\mathcal{A}}\Big|\Big|\frac{1}{n}\sum_{i=1}^{n} g_i(\theta_0)\mathbb{I}\{z_i\in A\}-\mathbb{E}[g(z,\theta_0)\mathbb{I}\{z\in A\}]\Big|\Big|>\epsilon\Big)\le 8d_g\Big(\frac{16end_gC_g}{(2v+1)\epsilon}\Big)^{2v+1}\exp\Big(-\frac{\epsilon^2 n}{128C_g^2d_g^2}\Big).$$

Then by Lemma D.1(i) and Lemma D.2(ii), the probability T_2^B exceeds $\epsilon/4$ is bounded above by

$$\begin{aligned}\mathbb{P}_n^{|}(T_2^B>\epsilon/4) &\le \mathbb{P}_n^{|}\Big(\sup_{A\in\mathcal{A}}\Big|\Big|\frac{1}{n}\sum_{i=1}^{n} g_i(\theta_0)\mathbb{I}\{z_i\in A\}-\mathbb{E}[g(z,\theta_0)\mathbb{I}\{z\in A\}]\Big|\Big|>\frac{\epsilon\delta_{\Omega,\min}}{4d_gC_g(1+\delta_{\Omega,\min})}\Big)\\ &\le 8d_g\Big(\frac{16end_gC_g}{(2v+1)\epsilon_b}\Big)^{2v+1}\exp\Big(-\frac{\epsilon_b^2 n}{128C_g^2d_g^2}\Big) \qquad (A.44)\end{aligned}$$

where $\epsilon_b=\epsilon\delta_{\Omega,\min}/4d_gC_g(1+\delta_{\Omega,\min})$.

For T_3^B, by a similar argument to that used to derive $(A.41)$,

$$\Big|\Big|\frac{1}{n}\sum_{i=1}^{n}(g_i(\hat\theta)-g_i(\theta_0))\Big|\Big|\le d_gC_G||\hat\theta-\theta_0||.$$

Hence, Lemma D.1(iv) and Lemma D.2(ii),

$$\begin{aligned}\mathbb{P}_n^{|}(T_3^B>\epsilon/4) &\le \mathbb{P}_n^{|}\Big(||\hat\theta-\theta_0||>\frac{\epsilon\delta_{\Omega,\min}}{4d_g^2C_GC_{B,\mathcal{A}}(1+\delta_{\Omega,\min})}\Big)\\ &\le \omega_1\exp\{-\omega_2\delta^2(\epsilon_c)n\}+\omega_3\exp\Big\{-\omega_4 n\Big\} \qquad (A.45)\end{aligned}$$

by Theorem 3.2 where $\epsilon_c=\epsilon\delta_{\Omega,\min}/4d_g^2C_GC_{B,\mathcal{A}}(1+\delta_{\Omega,\min})$, and the constants ω_1, ω_2, ω_3 and ω_4 are defined in $(A.28)$, $(A.29)$, $(A.33)$ and $(A.34)$.

For T_4^B, by Lemma D.1(iv) and Lemma D.2(ii),

$$\begin{aligned}\mathbb{P}_n^{|}(T_4^B > \epsilon/4) &\leq \mathbb{P}_n^{|}\Big(||g_n(\theta_0) - \mathbb{E}[g(z,\theta_0)]|| > \frac{\epsilon\delta_{\Omega,\min}}{4d_g C_{B,\mathcal{A}}(1+\delta_{\Omega,\min})}\Big) \\ &\leq 2d_g \exp\Big(\frac{-n\epsilon_d^2}{2d_g^2 C_g^2}\Big) \qquad (A.46)\end{aligned}$$

for $\epsilon_d = \epsilon\delta_{\Omega,\min}/4d_g C_{B,\mathcal{A}}(1+\delta_{\Omega,\min})$, where the second inequality follows from $(A.40)$.

The above arguments are conditional on the event $E_{\Omega,n}$. From $(A.35)$, the following conditional probability bound holds.

$$\begin{aligned}\mathbb{P}_n\big(|f_n(\hat{\theta},A)| > \epsilon \Big| E_{\Omega,n}\big) &\leq \min\Bigg\{\Big[\omega_1 \exp\{-\omega_2\delta^2(\epsilon^\star)n\} + \omega_3 \exp\Big\{-\omega_4 n\Big\}\Big] + \Big[2d_g \exp\Big(\frac{-n(\epsilon^{\star\star})^2}{2d_g^2 C_g^2}\Big)\Big] \\ &\quad , \Big[\omega_1 \exp\{-\omega_2\delta^2(\epsilon_a)n\} + \omega_3 \exp\Big\{-\omega_4 n\Big\}\Big] + \Big[8d_g\Big(\frac{16end_gC_g}{(2v+1)\epsilon_b}\Big)^{2v+1}\exp\Big(-\frac{\epsilon_b^2 n}{128C_g^2 d_g^2}\Big)\Big] \\ &\quad + \Big[\omega_1 \exp\{-\omega_2\delta^2(\epsilon_c)n\} + \omega_3 \exp\Big\{-\omega_4 n\Big\}\Big] + \Big[2d_g \exp\Big(\frac{-n\epsilon_d^2}{2d_g^2 C_g^2}\Big)\Big]\Bigg\} \qquad (A.47)\end{aligned}$$

where the constants ω_1, ω_2, ω_3 and ω_4 are defined in $(A.28)$, $(A.29)$, $(A.33)$ and $(A.34)$, and $\epsilon^\star = \epsilon\delta_{\Omega,\min}/2(1+\delta_{\Omega,\min})d_g C_g$, $\epsilon^{\star\star} = \epsilon\delta_{\Omega,\min}/2(1+\delta_{\Omega,\min})d_g C_g$, $\epsilon_a = \epsilon\delta_{\Omega,\min}/4d_g^2 C_G C_g(1+\delta_{\Omega,\min})$, $\epsilon_b = \epsilon\delta_{\Omega,\min}/4d_g C_g(1+\delta_{\Omega,\min})$, $\epsilon_c = \epsilon\delta_{\Omega,\min}/4d_g^2 C_G C_{B,\mathcal{A}}(1+\delta_{\Omega,\min})$ and $\epsilon_d = \epsilon\delta_{\Omega,\min}/4d_g C_{B,\mathcal{A}}(1+\delta_{\Omega,\min})$.

By similar arguments used for $(A.31)$, by Lemma B.1,

$$\begin{aligned}\mathbb{P}_n(|f_n(\hat{\theta},A)| > \epsilon) &\leq \mathbb{P}_n(\big\{|f_n(\hat{\theta},A)| > \epsilon\big\} \cap E_{\Omega,n}) + \mathbb{P}_n(E_{\Omega,n}^c) \\ &= \mathbb{P}_n(|f_n(\hat{\theta},A)| > \epsilon | E_{\Omega,n})\mathbb{P}_n(E_{\Omega,n}) + \mathbb{P}_n(E_{\Omega,n}^c) \\ &\leq \mathbb{P}_n(|f_n(\hat{\theta},A)| > \epsilon | E_{\Omega,n}) + \mathbb{P}_n(E_{\Omega,n}^c). \qquad (A.48)\end{aligned}$$

As before, by $(A.27)$ with $\kappa_3 = 1$

$$\mathbb{P}_n(E_{\Omega,n}^c) \leq \omega_3 \exp\Big\{-\omega_4 n\Big\}. \qquad (A.49)$$

Finally, by $(A.47)$, $(A.48)$ and $(A.49)$,

$$\begin{aligned}\mathbb{P}_n(|f_n(\hat{\theta},A)| > \epsilon) &\leq \min\Bigg\{\Big[\omega_1 \exp\{-\omega_2\delta^2(\epsilon^\star)n\} + \omega_3 \exp\Big\{-\omega_4 n\Big\}\Big] + \Big[2d_g \exp\Big(\frac{-n(\epsilon^{\star\star})^2}{2d_g^2 C_g^2}\Big)\Big] \\ &\quad , \Big[\omega_1 \exp\{-\omega_2\delta^2(\epsilon_a)n\} + \omega_3 \exp\Big\{-\omega_4 n\Big\}\Big] + \Big[8d_g\Big(\frac{16end_gC_g}{(2v+1)\epsilon_b}\Big)^{2v+1}\exp\Big(-\frac{\epsilon_b^2 n}{128C_g^2 d_g^2}\Big)\Big] \\ &\quad + \Big[\omega_1 \exp\{-\omega_2\delta^2(\epsilon_c)n\} + \omega_3 \exp\Big\{-\omega_4 n\Big\}\Big] + \Big[2d_g \exp\Big(\frac{-n\epsilon_d^2}{2d_g^2 C_g^2}\Big)\Big]\Bigg\} + \omega_3 \exp\Big\{-\omega_4 n\Big\}\end{aligned}$$

where the constants ω_1, ω_2, ω_3 and ω_4 are defined in $(A.28)$, $(A.29)$, $(A.33)$ and $(A.34)$, and $\epsilon^\star = \epsilon\delta_{\Omega,\min}/(2(1+\delta_{\Omega,\min})d_g C_g)$, $\epsilon^{\star\star} = \epsilon\delta_{\Omega,\min}/(2(1+\delta_{\Omega,\min})d_g C_g)$, $\epsilon_a = \epsilon\delta_{\Omega,\min}/(4d_g^2 C_G C_g(1+\delta_{\Omega,\min}))$, $\epsilon_b = \epsilon\delta_{\Omega,\min}/(4d_g C_g(1+\delta_{\Omega,\min}))$, $\epsilon_c = \epsilon\delta_{\Omega,\min}/(4d_g^2 C_G C_{B,\mathcal{A}}(1+\delta_{\Omega,\min}))$ and $\epsilon_d =$

$\epsilon\delta_{\Omega,\min}/(4d_gC_{B,\mathcal{A}}(1+\delta_{\Omega,\min}))$. □

Proof of Theorem 4.1

Part (i). Since $\hat{W}_n(\hat{A}) \geq \hat{W}_n(\tilde{A})$, as in Kitagawa and Tetenov [19], write

$$\begin{aligned} W(\tilde{A}) - W(\hat{A}) &\leq (W_n(\tilde{A}) - \hat{W}_n(\tilde{A})) - (W_n(\hat{A}) - \hat{W}_n(\hat{A})) \\ &\quad + (W(\tilde{A}) - W_n(\tilde{A})) - (W(\hat{A}) - W_n(\hat{A})) \\ &:= T_1 - T_2 + T_3 - T_4. \end{aligned}$$

For T_1, note that

$$\begin{aligned} W_n(\tilde{A}) - \hat{W}_n(\tilde{A}) &= \sum_{k=1}^{d_g} \Big(\frac{1}{n}\sum_{i=1}^{n} g^{(k)}(z_i,\theta_0)\mathbb{I}\{w_i \in \tilde{A}\}\Big)^2 - \Big(\frac{1}{n}\sum_{i=1}^{n} g^{(k)}(z_i,\hat{\theta})\mathbb{I}\{w_i \in \tilde{A}\}\Big)^2 \\ &= \sum_{k=1}^{d_g} \Big(\frac{1}{n}\sum_{i=1}^{n} [g^{(k)}(z_i,\theta_0) - g^{(k)}(z_i,\hat{\theta})]\mathbb{I}\{w_i \in \tilde{A}\}\Big) \\ &\quad \times \Big(\frac{1}{n}\sum_{i=1}^{n} [g^{(k)}(z_i,\theta_0) + g^{(k)}(z_i,\hat{\theta})]\mathbb{I}\{w_i \in \tilde{A}\}\Big) \end{aligned}$$

Under Assumption 2.1(ii) and CS,

$$\sup_{A\in\mathcal{A}} \Big|\frac{1}{n}\sum_{i=1}^{n} [g^{(k)}(z_i,\theta_0) + g^{(k)}(z_i,\hat{\theta})]\mathbb{I}\{w_i \in A\}\Big| \leq 2C_g.$$

By identical arguments used to establish $(A.2)$, there exist constants $\dot{C} > 0$ and $\dot{c} > 0$ such that

$$\mathbb{P}\Big(\sup_{A\in\mathcal{A}} \Big|\sum_{k=1}^{d_g} \Big(\frac{1}{n}\sum_{i=1}^{n} [g^{(k)}(z_i,\theta_0) - g^{(k)}(z_i,\hat{\theta})]\mathbb{I}\{w_i \in A\}\Big)\Big| > \epsilon\Big) \leq \dot{C}\exp(-\dot{c}n).$$

Thus, by CS, $|T_1| \leq o_p(1)$. Similarly, $|T_2| \leq o_p(1)$.

Similarly to T_1, for T_3, note that

$$\begin{aligned} |W(\tilde{A}) - W_n(\tilde{A})| &\leq \sum_{k=1}^{d_g} \Big|\frac{1}{n}\sum_{i=1}^{n} g^{(k)}(z_i,\theta_0)\mathbb{I}\{w_i \in \tilde{A}\} - \mathbb{E}[g^{(k)}(z,\theta_0)\mathbb{I}\{w \in \tilde{A}\}]\Big| \\ &\quad \times \Big|\frac{1}{n}\sum_{i=1}^{n} g^{(k)}(z_i,\theta_0)\mathbb{I}\{w_i \in \tilde{A}\} + \mathbb{E}[g^{(k)}(z,\theta_0)\mathbb{I}\{w \in \tilde{A}\}]\Big|. \end{aligned}$$

Under Assumptions 2.1(ii) and 2.2(viii), and CS,

$$\sup_{A\in\mathcal{A}}\Big|\frac{1}{n}\sum_{i=1}^{n} g^{(k)}(z_i,\theta_0)\mathbb{I}\{w_i\in A\}+\mathbb{E}[g^{(k)}(z,\theta_0)\mathbb{I}\{w\in A\}]\Big|\leq C_g+C_{B,\mathcal{A}}.$$

By Lemma 3.1, there exist constants $\ddot{C}>0$ and $\ddot{c}>0$ such that

$$\mathbb{P}\Big(\sup_{A\in\mathcal{A}}\Big|\frac{1}{n}\sum_{i=1}^{n} g^{(k)}(z_i,\theta_0)\mathbb{I}\{w_i\in A\}-\mathbb{E}[g^{(k)}(z,\theta_0)\mathbb{I}\{w\in A\}]\Big|>\epsilon\Big)\leq \ddot{C}\exp(-\ddot{c}n).$$

Thus, $|T_3|\leq o_p(1)$. Similarly, $|T_4|\leq o_p(1)$. Therefore, there exist constants $C_0>0$ and $c_0>0$ such that

$$\mathbb{P}(W(\tilde{A})-W(\hat{A})>\epsilon)\leq C_0\exp(-c_0 n).$$

By definition $W(\tilde{A})\geq W(\hat{A})\geq 0$, so that a bound for $|W(\hat{A})-W(\tilde{A})|$ is obtained from the above.

Part (ii). Let $Q_0(\mathcal{A})=\inf_{A\in\mathcal{A}}\{-W(A)\}$, $Q_{0,n}(\mathcal{A})=\inf_{A\in\mathcal{A}}\{-\hat{W}_n(A)\}$, $Q(A)=-W(A)-Q_0(\mathcal{A})$ and $Q_n(A)=-\hat{W}_n(A)-Q_{0,n}(\mathcal{A})$. Therefore, without loss of generality, $\tilde{A}=\arg\inf_{A\in\mathcal{A}}Q(A)$ and $\hat{A}=\arg\inf_{A\in\mathcal{A}}Q_n(A)$. Furthermore, $Q(A)\geq 0$ and $Q_n(A)\geq 0$ are equal to zero at $\tilde{A}$ and $\hat{A}$, respectively.

Define the ϵ-expansions of $\tilde{A}$ and $\hat{A}$ respectively as the classes of sets $\tilde{A}^\epsilon=\{w\in A:\ A\in\mathcal{A}$ and $d(w,\tilde{A})\leq\epsilon\}$ and $\hat{A}^\epsilon=\{w\in A:\ A\in\mathcal{A}$ and $d(w,\hat{A})\leq\epsilon\}$.

To fix ideas, let $\mathcal{A}$ be a class of sets formed by linear discrimination. Then,

$$\mathcal{A}=\{w\in\mathbb{R}^{d_w}:\ \alpha+w'\beta\geq 0,\ \alpha\in\mathbb{R},\ \beta\in\mathbb{R}^{d_w}\}.$$

A typical set $\bar{A}\in\mathcal{A}$, for some particular $\bar{\alpha}\in\mathbb{R}$, $\bar{\beta}\in\mathbb{R}^{d_w}$, is

$$\bar{A}=\{w\in\mathbb{R}^{d_w}:\ \bar{\alpha}+w'\bar{\beta}\geq 0\}.$$

If $\tilde{A}=\{w\in\mathbb{R}^{d_w}:\ \alpha_0+w'\beta_0\geq 0\}$ for some $\alpha_0\in\mathbb{R}$, $\beta_0\in\mathbb{R}^{d_w}$, then the ϵ-expansion of $\tilde{A}$ is

$$\begin{aligned}\tilde{A}^\epsilon \;=\;& \{w\in\mathbb{R}^{d_w}:\ \alpha+w'\beta\geq 0,\ |\alpha-\alpha_0|\leq\epsilon_1,\ ||\beta-\beta_0||\leq\epsilon_2,\\ & \text{where } \epsilon_1,\epsilon_2>0 \text{ allow } d(w,\tilde{A})\leq\epsilon\}.\end{aligned}$$

Consider the following ϵ-expansions of a set $A^\star$ for linear classifiers and rectangle classifiers.

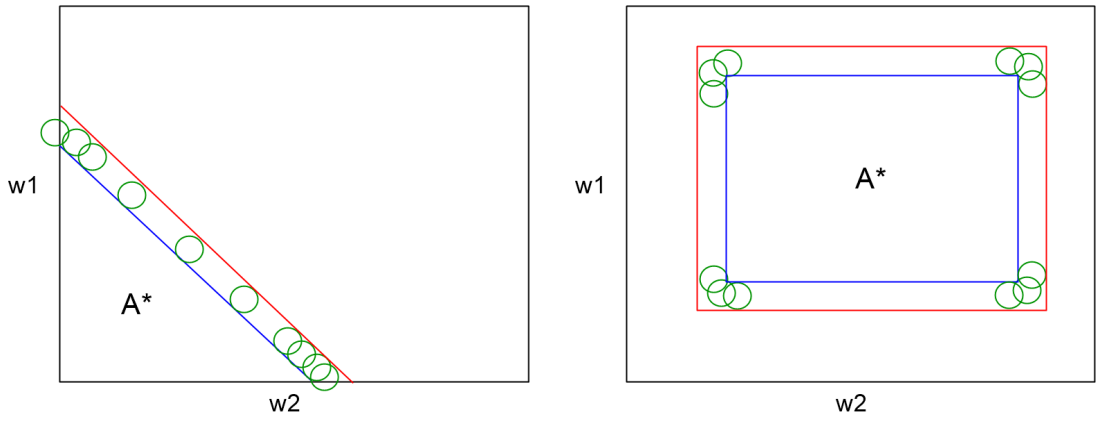


Figure A.1. Examples of ϵ-expansions of sets.

For $w = (w_1, w_2) \in \mathcal{W}^2$ where $\mathcal{W}$ is the support of each w_1 and w_2, consider the set $A^\star$ in blue. On the left $A^\star$ is formed by a linear classifier, and on the right $A^\star$ is formed by a rectangle classifier. The green balls represent a distance ϵ away from any point in that is contained in the set $A^\star$. An ϵ-expansion of $A^\star$, $A^{\star\epsilon}$, then extends the blue border of $A^\star$ and includes all points within the red border. As can be easily seen in the case of rectangle classifiers, this does not mean all points in $A^{\star\epsilon}$ are at most a distance ϵ away from $A^\star$, since some points in the corner of $A^{\star\epsilon}$ are slightly further away. However, by choosing sufficiently small balls ϵ, any point in $A^{\star\epsilon}$ can be arbitrarily close to any point in $A^\star$.

If $\mathcal{A}$ is a class of sets with VC dimension $v < \infty$, then $\tilde{A}^\epsilon$ is a class of sets with VC dimension at most v. To see this, let $\mathcal{A}$ shatter $\{w_{j_1}, ..., w_{j_v}\}$, $(j = 1, 2, ...)$. Then, unless $d(w_{j_l}, \tilde{A}) \leq \epsilon$, $(l = 1, ..., v)$, for some j, $\tilde{A}^\epsilon$ can only shatter less than v points in $\mathcal{W}$. Hence, the VC dimension of $\tilde{A}^\epsilon$ is at most v.

By Lemma 2.6.17(i) of van der Vaart and Wellner [35], p.147, the class of sets $(\tilde{A}^\epsilon)^c = \{A^c : A \in \tilde{A}^\epsilon\}$ has VC dimension at most v. Similarly, $\hat{A}^\epsilon$ and $(\hat{A}^\epsilon)^c$ are classes of sets with VC dimension at most v.

To prove $d_H(\hat{A}, \tilde{A}) = o_p(1)$, it is sufficient to show that, for any arbitrary $\epsilon > 0$, w.p.a.1 (a) $\sup_{w \in \hat{A}} d(w, \tilde{A}) \leq \epsilon$ and (b) $\sup_{w \in \tilde{A}} d(w, \hat{A}) \leq \epsilon$.

(a) $\sup_{w \in \hat{A}} d(w, \tilde{A}) \leq \epsilon$.

Let $\hat{A}_1 = \arg\inf_{A \in (\tilde{A}^\epsilon)^c} Q_n(A)$ and $\tilde{A}_1 = \arg\inf_{A \in (\tilde{A}^\epsilon)^c} Q(A)$. By definition, since $\tilde{A}$ is a unique minimiser of $Q(A)$ over $A \in \mathcal{A}$, $Q(\tilde{A}_1) \geq \delta(\epsilon)$ for some $\delta(\epsilon) > 0$. Also, by definition, $Q(\tilde{A}) = 0$. Hence, $Q(\hat{A}) - Q(\tilde{A}) = Q(\hat{A}) = W(\tilde{A}) - W(\hat{A}) = o_p(1)$, where the last equality follows by part (i). Therefore, $Q(\hat{A}) = o_p(1) < \delta(\epsilon) = Q(\tilde{A}_1)$ w.p.a.1. Since $\tilde{A}_1$ minimises $Q(A)$ over $A \in (\tilde{A}^\epsilon)^c$, it follows that $\hat{A} \in \tilde{A}^\epsilon$ w.p.a.1. That is, if $w \in \hat{A}$, then $w \in \tilde{A}^\epsilon$ w.p.a.1.

(b) $\sup_{w \in \tilde{A}} d(w, \hat{A}) \leq \epsilon$.

Here, $\hat{A}_2 = \arg\inf_{A \in (\hat{A}^\epsilon)^c} Q_n(A)$ and $\tilde{A}_2 = \arg\inf_{A \in (\hat{A}^\epsilon)^c} Q(A)$. By definition, since $\hat{A}$ is a unique minimiser of $Q_n(A)$ over $A \in \mathcal{A}$, $Q_n(\hat{A}_2) \geq \delta(\epsilon) > 0$. Also, $Q_n(\hat{A}) = 0$, so that

$Q_n(\tilde{A}) - Q_n(\hat{A}) = Q_n(\tilde{A}) = \hat{W}_n(\hat{A}) - \hat{W}_n(\tilde{A})$. Re-write,

$$\begin{aligned}\hat{W}_n(\hat{A}) - \hat{W}_n(\tilde{A}) &= (\hat{W}_n(\hat{A}) - W_n(\hat{A})) + (W_n(\hat{A}) - W(\hat{A})) + (W(\hat{A}) - W(\tilde{A})) \\ &\quad +(W(\tilde{A}) - W_n(\tilde{A})) + (W_n(\tilde{A}) - \hat{W}_n(\tilde{A})).\end{aligned}$$

Hence, $\hat{Q}_n(\tilde{A}) \leq 2\sup_{A\in\mathcal{A}} |\hat{W}_n(A) - W_n(A)| + 2\sup_{A\in\mathcal{A}} |W_n(A) - W(A)| + |W(\tilde{A}) - W(\hat{A})| \leq o_p(1)$, by similar arguments used for part (i). Thus, $\hat{Q}_n(\tilde{A}) = o_p(1) < \delta(\epsilon) \leq \hat{Q}_n(\hat{A}_2)$ w.p.a.1. Hence, w.p.a.1, $\tilde{A} \in \hat{A}^\epsilon$, that is, if $w \in \tilde{A}$, then $w \in \hat{A}^c$ w.p.a.1.

Combining (a) and (b), since the choice of $\epsilon > 0$ is arbitrary, the required result holds. □

Proof of Corollary 5.1

Let $\tilde{B}$ be the $d_g \times s$ matrix with the (k,j)-th entry given by $B^{(k)}(\tilde{A}_j) = \mathbb{E}[g^{(k)}(z,\theta_0)\mathbb{I}\{z \in \tilde{A}_j\}]$. From the Proof of Lemma 3.1 of RS,

$$\sqrt{n}\hat{\mu}_{s,n} = \sqrt{n}\hat{B}'_{s,n}\hat{\lambda} + o_p(1).$$

For the $d_g \times s$ matrices $\hat{B}_{s,n}$ and $\tilde{B}$,

$$\begin{aligned}||\hat{B}_{s,n} - \tilde{B}|| &\leq ||\hat{B}_{s,n} - \tilde{B}||_1 \\ &= \max_{j=1,...,s} \sum_{k=1}^{d_g} |\hat{B}_n^{(k)}(\hat{A}_j) - B^{(k)}(\tilde{A}_j)|\end{aligned}$$

By T, for any $k \in \{1,...,d_g\}$, $j \in \{1,...,s\}$,

$$\begin{aligned}|\hat{B}_n^{(k)}(\hat{A}_j) - B^{(k)}(\tilde{A}_j)| &\leq |\hat{B}_n^{(k)}(\hat{A}_j) - B_n^{(k)}(\hat{A}_j)| + |B_n^{(k)}(\hat{A}_j) - B^{(k)}(\hat{A}_j)| \\ &\quad +|B^{(k)}(\hat{A}_j) - B^{(k)}(\tilde{A}_j)| \\ &:= (K1) + (K2) + (K3). \qquad \text{(A.50)}\end{aligned}$$

For $(K1)$, by T, for some $\bar{\theta}$ on a line segment joining $\hat{\theta}$ and θ_0,

$$\begin{aligned}\left|\left|\frac{1}{n}\sum_{i=1}^n [g^{(k)}(z_i\hat{\theta}) - g^{(k)}(z_i\theta_0)]\mathbb{I}\{z_i \in \hat{A}_j\}\right|\right| &\leq \frac{1}{n}\sum_{i=1}^n ||g^{(k)}(z_i,\hat{\theta}) - g^{(k)}(z_i,\theta_0)|| \\ &\leq \frac{1}{n}\sum_{i=1}^n ||G(z_i,\bar{\theta})|| \times ||\hat{\theta} - \theta_0|| \\ &\leq C_G||\hat{\theta} - \theta_0|| \\ &= o_p(1),\end{aligned}$$

where the second inequality follows by a Taylor expansion, the third from Assumption 2.1(iii) and the fourth from the consistency of the GEL estimator $\hat{\theta}$ for θ_0.

For $(K2)$,

$$\begin{aligned} |B_n^{(k)}(\hat{A}_j) - B^{(k)}(\hat{A}_j)| &\leq \sup_{A \in \mathcal{A} \setminus \hat{A}_1, ..., \hat{A}_{j-1}} |B_n^{(k)}(A) - B^{(k)}(A)| \\ &\leq o_p(1), \end{aligned}$$

by Lemmata C.1 and C.5.

For $(K3)$, write

$$\begin{aligned} B(\hat{A}) &= \mathbb{E}[g^{(k)}(z, \theta_0) \mathbb{I}\{z \in \hat{A}_j\}] \\ &= \int_{\mathcal{Z}} g^{(k)}(z, \theta_0) \mathbb{I}\{z \in \hat{A}_j\} F(dz), \end{aligned}$$

where Lebesgue integration is taken with respect to the distribution $F(z)$ of z.

The indicator function can be approximated by a smooth function, in particular, by convolution of the indicator function with a smooth bump function, as in the literature of mollifiers. Let $\psi : \mathbb{R} \to \mathbb{R}$ be the function $\Psi(u) = c \exp(-1/(1-u^2))$ for $u \in (-1, 1)$, and $\Psi(u) = 0$ otherwise, for some positive constant c. Also, for $y = (y_1, ..., y_{d_z}) \in \mathbb{R}^{d_z}$ construct $\Phi(y) = c\Psi(y_1)\Psi(y_2)...\Psi(y_{d_z})$, where c ensures normalisation such that $\int_{-1}^{1} \Phi(y) dy = 1$. For $\epsilon > 0$, define $\Phi_\epsilon(y) = \Phi(y/\epsilon)/\epsilon$.

The indicator function is a function in L^p space ($1 \leq p < \infty$) since for any set $A \subset \mathbb{R}^{d_z}$, $||\mathbb{I}_A||_p = \left(\int_{\mathcal{Z}} |\mathbb{I}_A|^p dF \right)^{1/p} \leq \left(\int_{\mathcal{Z}} dF \right)^{1/p} < \infty$. Therefore, applying results of convolution approximation of functions in L^p space, for example, Theorems 9.5 and 9.10 of Rudin [31], $||\mathbb{I}_{\hat{A}_j} * \Phi_\epsilon - \mathbb{I}_{\hat{A}_j}||_p \to 0$ as $\epsilon \to 0$. Similarly, $||\mathbb{I}_{\tilde{A}_j} * \Phi_\epsilon - \mathbb{I}_{\tilde{A}_j}||_p \to 0$ as $\epsilon \to 0$.

Therefore,

$$\begin{aligned} \int_{\mathcal{Z}} g^{(k)}(z, \theta_0) \mathbb{I}\{z \in \hat{A}_j\} dF(z) &= \int_{\mathcal{Z}} \left[\lim_{\epsilon \to 0} (\mathbb{I}_{\hat{A}_j} * \Phi_\epsilon)(z) \right] g^{(k)}(z, \theta_0) F(dz) + o_p(1) \\ &\int_{\mathcal{Z}} \left[\lim_{\epsilon \to 0} (\mathbb{I}_{\tilde{A}_j} * \Phi_\epsilon)(z) \right] g^{(k)}(z, \theta_0) F(dz) + o_p(1) \\ &\int_{\mathcal{Z}} g^{(k)}(z, \theta_0) \mathbb{I}\{z \in \tilde{A}_j\} F(dz) + o_p(1). \end{aligned}$$

where the first equality follows from a convolution approximation, the second equality follows from Hausdorff convergence of sets $\hat{A}_j$ and $\tilde{A}_j$, part (ii) of Theorem 4.1, and the third equality follows from a convolution approximation. Hence, $(K3) = o_p(1)$.

Thus, from $(A.50)$, $|\hat{B}_n^{(k)}(\hat{A}_j) - B^{(k)}(\tilde{A}_j)| \leq o_p(1)$. Therefore, by T,

$$\begin{aligned} \left|\left| \hat{B}_{s,n} - \tilde{B} \right|\right| &\leq \sum_{k=1}^{d_g} \left| \hat{B}_n^{(k)}(\hat{A}_j) - B^{(k)}(\tilde{A}_j) \right| \\ &= d_g \times o_p(1) = o_p(1). \end{aligned} \tag{A.51}$$

Therefore,

$$\sqrt{n}\hat{B}'_{s,n}\hat{\lambda} = \sqrt{n}\tilde{B}'\hat{\lambda} + o_p(1).$$

The rest of the proof follows that of Theorem 3.2 of RS. Define $P = \Omega^{-1} - \Omega^{-1}G(G'\Omega^{-1}G)^{-1}G'\Omega^{-1}$. If $\text{rank}(\tilde{B}) = d_g$, then $\tilde{B}'(\tilde{B}\tilde{B}')^{-1}\Omega(\tilde{B}\tilde{B}')^{-1}\tilde{B}$ is a generalised-inverse of $\tilde{B}'P\tilde{B}$ as $P\Omega P = P$. Furthermore by $(A.51)$ and $\hat{\Omega}$ a consistent estimator of Ω, by the continuous mapping theorem, $\hat{B}'_{s,n}(\hat{B}_{s,n}\hat{B}'_{s,n})^{-1}\hat{\Omega}(\hat{B}_{s,n}\hat{B}'_{s,n})^{-1}\hat{B}_{s,n} \xrightarrow{p} \tilde{B}'(\tilde{B}\tilde{B}')^{-1}\Omega(\tilde{B}\tilde{B}')^{-1}\tilde{B}$. Using the expansion $\sqrt{n}\hat{\lambda} = -P\sqrt{n}g_n(\theta_0) + o_p(1)$ (cf. Newey and Smith [23], proof of Theorem 3.2, p. 240),

$$\begin{aligned} T_n &= n\hat{\mu}'_{s,n}\hat{B}'_{s,n}(\hat{B}_{s,n}\hat{B}'_{s,n})^{-1}\hat{\Omega}(\hat{B}_{s,n}\hat{B}'_{s,n})^{-1}\hat{B}_{s,n}\hat{\mu}_{s,n} + o_p(1) \\ &= n g_n(\theta_0)'P\Omega P g_n(\theta_0) + o_p(1), \end{aligned}$$

is asymptotically equivalent to the Lagrange multiplier test for overidentifying moment conditions which has an asymptotic $\chi^2_{d_g - d_\theta}$ distribution. The required result follows. □

B Probability Bounds

B.1 Probability of Events

The following lemma will be used to provide probability bounds on individual events.

Lemma B.1. *Consider events $E_1, ..., E_S$. For all $S \geq 2$,*

$$\mathbb{P}(E_1) \leq \mathbb{P}\Big(\bigcap_{s=1}^{S} E_s\Big) + \sum_{s\neq 1}^{S} \mathbb{P}(E_s^c)$$

where $\mathbb{P}(E_s^c) = 1 - \mathbb{P}(E_s)$ for all $s \in \{1, ..., S\}$.

Proof. (By induction). Consider the base case $S = 2$. Then,

$$\begin{aligned} \mathbb{P}(E_1 \cap E_2) &= \mathbb{P}(E_1) + \mathbb{P}(E_2) - \mathbb{P}(E_1 \cup E_2) \\ &\geq \mathbb{P}(E_1) + \mathbb{P}(E_2) - 1 \end{aligned}$$

or, $\mathbb{P}(E_1) \leq \mathbb{P}(E_1 \cap E_2) + \mathbb{P}(E_2^c)$. For the inductive step, assume the statement in the Lemma holds with $S = k$. Then,

$$\mathbb{P}\Big(\bigcap_{s=1}^{k} E_s\Big) \geq \mathbb{P}(E_1) - \sum_{s\neq 1}^{k} \mathbb{P}(E_s^c). \tag{B.1}$$

Then,

$$\begin{aligned}
\mathbb{P}\Big(\bigcap_{s=1}^{k+1} E_s\Big) &= \mathbb{P}(E_{k+1}) + \mathbb{P}\Big(\bigcap_{s=1}^{k} E_s\Big) - \mathbb{P}\Big(E_{k+1} \cup \bigcap_{s=1}^{k} E_s\Big) \\
&\geq \mathbb{P}(E_{k+1}) + \mathbb{P}\Big(\bigcap_{s=1}^{k} E_s\Big) - 1 \\
&= \mathbb{P}\Big(\bigcap_{s=1}^{k} E_s\Big) - \mathbb{P}(E_{k+1}^c) \\
&\geq \mathbb{P}(E_1) - \sum_{s\neq 1}^{k+1} \mathbb{P}(E_s^c)
\end{aligned}$$

where the last line follows from $(B.1)$. Therefore if the statement in the lemma holds for any $S = k$, then it also holds for $S = k + 1$. Since the statement holds for $S = 2$, it follows that the statement holds for all $S \geq 2$. □

Lemma B.2. *Consider random quantities $X_1, ..., X_S$. For all $S \geq 2$ and $\epsilon > 0$,*

$$\mathbb{P}(\sum_{s=1}^{S} X_s > \epsilon) \leq \sum_{s=1}^{S} \mathbb{P}\Big(X_s > \frac{\epsilon}{S}\Big).$$

Proof. Note that if the event $\big\{\sum_{s=1}^{S} X_s > \epsilon\big\}$ holds then for at least one $s \in \{1, ..., S\}$ the event $\big\{X_s > \frac{\epsilon}{S}\big\}$ must hold. Therefore, by the union bound,

$$\begin{aligned}
\mathbb{P}\Big(\sum_{s=1}^{S} X_s > \epsilon\Big) &\leq \mathbb{P}\Big(\bigcup_{s=1} \{X_s > \frac{\epsilon}{S}\}\Big) \\
&\leq \sum_{s=1}^{S} \mathbb{P}\Big(X_s > \frac{\epsilon}{S}\Big).
\end{aligned}$$

□

B.2 Large Deviation Bounds

The following results on large deviation bounds will be utilised.

Lemma B.3 (Hoeffding's inequality)**.** *Let $f_1, ..., f_n$ be real-valued, zero-mean, independent functions defined on $\mathcal{Z}$ with $f_i : \mathcal{Z} \to [a_i, b_i]$ for $i = 1, ..., n$, where a_i, b_i are real numbers satisfying $a_i < b_i$. Then,*

$$\mathbb{P}\Big\{\Big|\sum_{i=1}^{n} f_i(z)\Big| > \epsilon\Big\} \leq 2exp\Big(-\frac{2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\Big).$$

Lemma B.4 provides bounds on the GEL Lagrange multiplier $\hat{\lambda}(\theta)$ away from $\lambda(\theta)$ for any fixed $\theta \in \Theta$, and the result follows from Theorem 2.1 of Inglot and Kallenberg [18] (also see Proof of Theorem 2.1(c) of Otsu [26], p.328).

Lemma B.4 (GEL Lagrange multipliers)**.** *Fix $\theta \in \Theta$. Under Assumptions 2.1-2.3, there exist positive constants c_λ and C_λ such that for any $n \in \mathbb{N}$,*

$$\mathbb{P}_n(||\hat{\lambda}(\theta) - \lambda(\theta)|| > \epsilon) \leq C_\lambda \exp(-c_\lambda n).$$

Lemma B.5 provides large deviation bounds on errors of GEL estimators. The result follows from Theorem 2.1(c) of Otsu [26].

Lemma B.5 (GEL estimators (Theorem 2.1(c) of Otsu [26]))**.** *Under Assumptions 2.1-2.3, there exist positive constants c_θ and C_θ such that for any $n \in \mathbb{N}$,*

$$\mathbb{P}_n(||\hat{\theta} - \theta_0|| > \epsilon) \leq C_\theta \exp\{-c_\theta n\}.$$

C Vapnik Chervonenkis Bounds

Definition. (Subgraph). The subgraph of a real-valued function $f : \mathcal{Z} \to \mathbb{R}$ is the subset of $\mathcal{Z} \times \mathbb{R}$ given by

$$SG(f) := \{(z,t) : 0 \leq t < f(z) \text{ or } f(z) < t < 0\}.$$

Let $\mathbb{I}_A(z) := \mathbb{I}\{z \in A\}$, and $\mathcal{I}$ be the class of indicator functions, $\mathcal{I} := \{\mathbb{I}_A(z),\ A \in \mathcal{A}\}$.

The following lemma follows from well known properties of VC classes. See for example, Lemma 2.6.18(vi) (p.147) of van der Vaart and Wellner [35], and Lemma 9.8 (p.154) and Lemma 9.9(vi) (p.155) of Kosorok [22]. For completeness, the proof is given below.

Lemma C.1. *Let $\mathcal{A}$ be a VC class of subsets of $\mathcal{Z}$ with VC dimension $v < \infty$. Let $h : \mathcal{Z} \to \mathbb{R}$ be a known function. Consider the set of functions from $\mathcal{Z}$ to $\mathbb{R}$,*

$$\mathcal{F} := \{f_A(z) = h(z) \cdot \mathbb{I}_A(z) \ : A \in \mathcal{A}\}.$$

Then $\mathcal{F}$ is a VC-subgraph class of functions with VC dimension $v_\mathcal{F} \leq 2v + 1$.

Proof. The proof verifies the following statements
(1) $\mathcal{I}$ is a VC-subgraph class of functions on $\mathcal{Z}$ with VC dimension v.
(2) Given (1), $\mathcal{F}$ is a VC-subgraph class of functions on $\mathcal{Z}$ with VC dimension $v_\mathcal{F} \leq 2v + 1$.

Let $SG(\mathcal{I})$ be the collection of subgraphs of the class of functions $\mathcal{I}$; $SG(\mathcal{I}) := \{SG(\mathbb{I}_A) : A \in \mathcal{A}\} = \{(z,t) : t < \mathbb{I}_A(z),\ A \in \mathcal{A}\}$. For any $t < 0$, $\mathbb{I}_A(z) > t$ for any z and A. No collection

$\{(z_1, t_1), ..., (z_k, t_k)\}$ can be shattered by $\mathcal{I}$ if any of the $t_j < 0$, $1 \leq j \leq k$. Similarly, no collection $\{(z_1, t_1), ..., (z_k, t_k)\}$ can be shattered by $\mathcal{I}$ if any of the $t_j \geq 1$, $1 \leq j \leq k$. Now note that the collection $\{(z_1, t_1), ..., (z_k, t_k)\}$ is only shattered if $\{(z_1, 0), ..., (z_k, 0)\}$ can be shattered. This occurs if and only if $\{z_1, ..., z_k\}$ is shattered by $\mathcal{A}$. Therefore $VC(\mathcal{I}) \leq VC(\mathcal{A}) = v$. This shows (1).

Express $\mathcal{F}$ as $\mathcal{I} \cdot h = \{h \cdot \mathbb{I}_A,\, A \in \mathcal{A}\}$. For any $\mathbb{I}_A \in \mathcal{I}$, the subgraph of $h \cdot \mathbb{I}_A$ is the union of the sets

$$\begin{aligned} \mathcal{F}_A^+ &:= \{(z,t) : t < h(z)\mathbb{I}_A(z),\, h(z) > 0\} \\ \mathcal{F}_A^- &:= \{(z,t) : t < h(z)\mathbb{I}_A(z),\, h(z) < 0\} \\ \mathcal{F}^0 &:= \{(z,t) : t < 0,\, h(z) = 0\}. \end{aligned}$$

Denote $\mathcal{F}^+ := \{\mathcal{F}_A^+ : A \in \mathcal{A}\}$. Consider $\mathcal{F}^+$ on $(\mathcal{Z} \cap \{h > 0\}) \times \mathbb{R}$. For any arbitrary $z \in \mathcal{Z}$ such that $h(z) > 0$, the subset $\{(z_1, t_1), ..., (z_k, t_k)\}$ is shattered by $\mathcal{F}^+$ if and only if the subset $\{(z_1, t_1/h(z_1)), ..., (z_k, t_k/h(z_k))$ is shattered by subgraphs of $\mathcal{I}$. Using (1), the VC dimension of $\mathcal{F}^+$ on this subset is $\leq v$. Similarly, the VC dimension of $\mathcal{F}^- := \{\mathcal{F}_A^- : A \in \mathcal{A}\}$ on $(\mathcal{Z} \cap \{h < 0\}) \times \mathbb{R}$ is $\leq v$. $\mathcal{F}^0$ on $(\mathcal{Z} \cap \{h = 0\}) \times \mathbb{R}$ cannot shatter two arbitrary points and therefore the VC dimension of $\mathcal{F}^0$ is 1.

Finally, use the result that if $\mathcal{Z}$ is the union of finitely disjoint sets $\mathcal{Z}_i$ and that and for each i, if $\mathcal{F}_i$ is a VC-subgraph class of functions on $\mathcal{Z}_i$ with VC dimension v_i, then $\cup\mathcal{F}_i$ is a VC-subgraph class of functions on $\cup\mathcal{Z}_i$ with VC dimension $\sum_i v_i$. (see proof of Lemma 2.6.18(vi) of van der Vaart and Wellner [35], p.148 and p.152). This shows (2). □

The concept of covering numbers is used in the proof of the VC inequality (see van der Vaart and Wellner [35] for a detailed treatment).

Definition. (Covering number). Let $\mathcal{F}$ be a class of functions from $\mathcal{Z} \times \Theta$ to $\mathbb{R}$. Let d be some metric. For $\varepsilon > 0$, $\mathcal{F}_\varepsilon \subset \mathcal{Z}$ is an ϵ-cover for $\mathcal{F}$ if $\mathcal{F}_\varepsilon \subset \mathcal{F}$ and for all $f \in \mathcal{F}$, there exists $q \in \mathcal{F}_\varepsilon$ such that $d(f, q) < \varepsilon$. If there is a finite cover for $\mathcal{F}$, the ε-covering number of $\mathcal{F}$ with respect to metric d, $\mathcal{N}(\varepsilon, \mathcal{F}, d)$, is the minimum cardinality of a ε-cover for $\mathcal{F}$.

A related quantity is the uniform covering number.

Definition. (Uniform covering number). Given a sequence $z = (z_1, ..., z_n) \in \mathcal{Z}^n$, let $F_{|z}$ be the subset of $\mathbb{R}^n$ given by $\mathcal{F}_{|z} = \{(f(z_1), ..., f(z_n)) : f \in \mathcal{F}\}$. For $\varepsilon > 0$, the uniform covering number $\mathcal{N}_d(\varepsilon, \mathcal{F}, n)$ is the maximum covering number $\mathcal{N}(\varepsilon, \mathcal{F}_{|z}, d)$ over all $z \in \mathcal{Z}^n$. i.e. $\mathcal{N}_d(\varepsilon, \mathcal{F}, n) = \max\{\mathcal{N}(\varepsilon, \mathcal{F}_{|z}, d) : z \in \mathcal{Z}^n\}$.[2]

[2] $\mathcal{N}_d(\varepsilon, \mathcal{F}, n)$ can be interpreted as a measure of the richness of the class $\mathcal{F}$ at the scale ε (Anthony and Bartlett [2]).

Let d_∞ be the metric on $\mathbb{R}^n$ defined by $d_\infty((x_1,...,x_n),(y_1,...,y_n)) = \max_i|x_i - y_i|$. The following result is from Anthony and Bartlett [2], p.167.

Lemma C.2. *Let $\mathcal{F} : \mathcal{Z} \times \Theta \to [-C_f, C_f]$ be a set of real-valued functions. Suppose $\mathcal{F}$ is a VC subgraph with VC dimension v. For all $n \geq v$,*

$$\mathcal{N}_\infty(\varepsilon, \mathcal{F}, n) \leq \Big(\frac{2enC_f}{v\varepsilon}\Big)^v.$$

The derivation of VC inequalities rely on symmetrisation arguments (Dudley [7]) which facilitate large deviation bounds involving expectations of functions. Lemma C.3 and Lemma C.4 below are standard results in empirical process theory. For example, see Theorems 10 and 11 of Anthony [1], and Theorem 1 of Bartlett and Lugosi [4].

Lemma C.3. (Symmetrisation I). *Let $z = (z_1,...,z_n) \in \mathcal{Z}^n$, and let $z' = (z'_1,...,z'_n) \in \mathcal{Z}^n$ be a second independent sample of length n. For some function $h : \mathcal{Z} \to \mathbb{R}$, define $f(z;A) = h(z)\mathbb{I}\{z \in A\}$, $f_n(A) = \sum_{i=1}^n f(z_i,A)/n$, and $f'_n(A) = \sum_{i=1}^n f(z'_i,A)/n$. Let $\sup_{z\in\mathcal{Z}}\sup_{A\in\mathcal{A}}|f(z,A)| \leq C_f$. Then for all $n > 8C_f^2/\delta^2$,*

$$\mathbb{P}_n\Big(\sup_{A\in\mathcal{A}}\Big|f_n(A) - \mathbb{E}f(z;A)\Big| > \delta\Big) \leq 2 \times \mathbb{P}_{2n}\Big(\sup_{A\in\mathcal{A}}\Big|f_n(A) - f'_n(A)\Big| > \frac{\delta}{2}\Big)$$

where the probability on the RHS is with respect to the product measure over $\mathcal{Z}^{2n}$.

Proof. First note that $\forall z_i, z'_i \in \mathcal{Z}^{2n}$, $\mathbb{E}[f(z_i;A)] = \mathbb{E}[f(z'_i;A)] = \mathbb{E}f(z,A)$. Let $A^\star \in \mathcal{A}$ be a set for which $|f_n(A^\star) - \mathbb{E}[f(z;A^\star)]| > \delta$ if such a set exists. Let $A^\star$ be a fixed set in $\mathcal{A}$ otherwise. By definition,

$$\mathbb{P}_{2n}\Big(\sup_{A\in\mathcal{A}}\Big|f_n(A) - f'_n(A)\Big| > \frac{\delta}{2}\Big) \geq \mathbb{P}_{2n}\Big(\big|f_n(A^\star) - f'_n(A^\star)\big| > \frac{\delta}{2}\Big).$$

By T,

$$|\mathbb{E}f(z,A^\star) - f_n(A^\star)| \leq |\mathbb{E}f(z,A^\star) - f'_n(A^\star)| + |f'_n(A^\star) - f_n(A^\star)|.$$

Since $|\mathbb{E}f(z,A^\star) - f_n(A^\star)| > \delta$, if $|\mathbb{E}f(z,A^\star) - f'_n(A^\star)| \leq \delta/2$, then $|f'_n(A^\star) - f_n(A^\star)| > \delta/2$. Therefore

$$\begin{aligned}
\mathbb{P}_{2n}\big(|f'_n(A^\star) - f_n(A^\star)| > \delta/2\big) &\geq \mathbb{P}_{2n}\big(|\mathbb{E}f(z,A^\star) - f_n(A^\star)| > \delta,\, |\mathbb{E}f(z,A^\star) - f'_n(A^\star)| \leq \delta/2\big) \\
&= \mathbb{E}_{2n}\big(\mathbb{I}\{|\mathbb{E}f(z,A^\star) - f_n(A^\star)| > \delta\}\mathbb{I}\{|\mathbb{E}f(z,A^\star) - f'_n(A^\star)| \leq \delta/2\}\big) \\
&= \mathbb{E}_{2n}\big[\mathbb{I}\{|\mathbb{E}f(z,A^\star) - f_n(A^\star)| > \delta\}\mathbb{P}_n\big(|\mathbb{E}f(z,A^\star) - f'_n(A^\star)| \leq \delta/2\big|z\big)\big]
\end{aligned}$$

where the second line follows from $\mathbb{E}[\mathbb{I}\{A\}\mathbb{I}\{B\}] = \mathbb{P}(A \cap B)$, and the third follows from the law of iterated expectations.

Popoviciu's inequality on variances states for a random variable X such that $m \leq X \leq M$, then $\text{var}(X) \leq (M-m)^2/4$. Therefore for $-C_f \leq f(z_i, A) \leq C_f$, $\text{var}(f(z_i, A)) \leq C_f^2$ for any $z_i \in \mathcal{Z}^{2n}$. Now

$$\begin{aligned}
\mathbb{P}_n\big(|f_n'(A^\star) - \mathbb{E}f(z, A^\star)| > \delta/2 \big| z_1, ..., z_n\big) &= \mathbb{P}_n\Big(\Big|\sum_{i=1}^n \big(f(z_i', A^\star) - \mathbb{E}f(z, A^\star)\big)\Big| > n\delta/2 \big| z_1, .., z_n\Big) \\
&\leq \text{var}\big(\sum_{i=1}^n f(z_i, A^\star)\big)/(n^2\delta^2/4) \\
&\leq nC_f^2/(n^2\delta^2/4) \\
&= 4C_f^2/n\delta^2
\end{aligned}$$

where the second line follows from Chebyshev's inequality. Then for $n > 8C_f^2/\delta^2$,

$$\mathbb{P}_n\big(|f_n'(A^\star) - \mathbb{E}f(z, A^\star)| \leq \delta/2 \big| z_1, ..., z_n\big) \geq 1 - \frac{4C_f^2}{n\delta^2} \geq \frac{1}{2}.$$

Overall,

$$\begin{aligned}
\mathbb{P}_{2n}\Big(\sup_{A\in\mathcal{A}} |f_n(A) - f_n'(A)| > \delta/2\Big) &\geq \frac{1}{2}\mathbb{P}_n\Big(|f_n(A^\star) - \mathbb{E}f(z, A^\star)| > \delta\Big) \\
&= \frac{1}{2}\mathbb{P}_n\Big(\sup_{A\in\mathcal{A}} |f_n(A) - \mathbb{E}f(z, A)| > \delta\Big).
\end{aligned}$$

□

Lemma C.4. (Symmetrisation II). *Let $\sigma_1, ..., \sigma_n$ be i.i.d. rademacher variables, independent of $z = z_1, ..., z_n$ and $z' = z_1', ..., z_n'$. For real valued functions f, define $f_n(z) = n^{-1}\sum_{i=1}^n f(z_i)$ and $f_n(z') = n^{-1}\sum_{i=1}^n f(z_i')$. Then,*

$$\mathbb{P}_{2n}\Big(\sup_{f\in\mathcal{F}} |f_n(z) - f_n'(z)| > \delta/2\Big) \leq 2 \times \mathbb{P}_n\Big(\sup_{f\in\mathcal{F}} \Big|\frac{1}{n}\sum_{i=1}^n \sigma_i f(z_i)\Big| > \frac{\delta}{4}\Big).$$

Proof. Due to exchangeability of $f(z_i)$ and $f(z_i')$, $f(z_i) - f(z_i')$ is a symmetric random variable.

I.e. $\mathbb{P}_{2n}(f(z_i) - f(z_i') \leq \epsilon) = \mathbb{P}_{2n}(f(z_i') - f(z_i) \leq \epsilon)$. Note that

$$\begin{aligned}
\mathbb{P}_{2n}(\sigma_i(f(z_i) - f(z_i')) \leq \epsilon) &= \mathbb{P}_{2n}(\sigma_i(f(z_i) - f(z_i')) \leq \epsilon | \sigma_i = 1)\mathbb{P}(\sigma_i = 1) \\
&\quad + \mathbb{P}_{2n}(\sigma_i(f(z_i) - f(z_i')) \leq \epsilon | \sigma_i = -1)\mathbb{P}(\sigma_i = -1) \\
&= \mathbb{P}_{2n}(f(z_i) - f(z_i') \leq \epsilon)\mathbb{P}(\sigma_i = 1) \\
&\quad + \mathbb{P}_{2n}(f(z_i') - f(z_i) \leq \epsilon)\mathbb{P}(\sigma_i = -1) \\
&= \mathbb{P}_{2n}(f(z_i) - f(z_i') \leq \epsilon)(\mathbb{P}(\sigma_i = 1) + \mathbb{P}(\sigma_i = -1)) \\
&= \mathbb{P}_{2n}(f(z_i) - f(z_i') \leq \epsilon).
\end{aligned}$$

Therefore the distribution of $f(z_i) - f(z_i')$ is the same as the distribution of $\sigma_i(f(z_i) - f(z_i'))$. Then

$$\begin{aligned}
\mathbb{P}_{2n}\Big(\sup_{f\in\mathcal{F}}|f_n(z) - f_n'(z)| > \delta/2\Big) &= \mathbb{P}_{2n}\Big(\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^n \sigma_i(f(z_i) - f(z_i'))\Big| > \frac{\delta}{2}\Big) \\
&\leq \mathbb{P}_{2n}\Big(\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^n \sigma_i f(z_i)\Big| + \Big|\frac{1}{n}\sum_{i=1}^n \sigma_i f(z_i)\Big| > \frac{\delta}{4} + \frac{\delta}{4}\Big) \\
&\leq \mathbb{P}_{n}\Big(\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^n \sigma_i f(z_i)\Big| > \frac{\delta}{4}\Big) + \mathbb{P}_{n}\Big(\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^n \sigma_i f(z_i')\Big| > \frac{\delta}{4}\Big) \\
&= 2 \times \mathbb{P}_{n}\Big(\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^n \sigma_i f(z_i)\Big| > \frac{\delta}{4}\Big)
\end{aligned}$$

where the final probability is taken over $\sigma_1, ..., \sigma_n$ and $z_1, ..., z_n$. □

Lemma C.5. (VC Inequality) *Under Assumptions 2.1-2.3 and 3.1, for all $n > \max\{2v + 1, 8C_g^2/\delta^2\}$ and $\delta > 0$, if $\mathcal{A}$ is a class of subsets of $\mathcal{Z}$ with VC dimension $v < \infty$, then for the function $f : \mathcal{Z} \times \Theta \times \mathcal{A} \to \mathbb{R}$, $f(z, A) = g^{(k)}(z, \theta_0)\mathbb{I}\{z \in A\}$ $(k = 1, ..., d_g)$ and $f_n(A) = \sum_{i=1}^n f(z_i, A)/n$,*

$$\mathbb{P}_n\Big(\sup_{A\in\mathcal{A}}\big|f_n(A) - \mathbb{E}f(z, A)\big| > \delta\Big) \leq 8\Big(\frac{16enC_g}{(2v+1)\delta}\Big)^{2v+1}\exp\Big(-\frac{\delta^2 n}{128C_g^2}\Big).$$

Proof. Let $z = (z_1, ..., z_n) \in \mathcal{Z}^n$, and let $z' = (z_1', ..., z_n') \in \mathcal{Z}^n$ be a second independent sample of length n. Let $f_n'(A) = n^{-1}\sum_{i=1}^n f(z_i', A)$. And note that $|f(z_i, A)| \leq C_g$ for all $z_i \in \mathcal{Z}^{2n}$. By Lemma C.3, $\forall n > 8C_g^2/\delta^2$,

$$\mathbb{P}_n\Big(\sup_{A\in\mathcal{A}}\Big|f_n(A) - \mathbb{E}f(z, A)\Big| > \delta\Big) \leq 2 \times \mathbb{P}_{2n}\Big(\sup_{A\in\mathcal{A}}\Big|f_n(A) - f_n'(A)\Big| > \frac{\delta}{2}\Big)$$

where the probability on the RHS is with respect to the product measure over $\mathcal{Z}^{2n}$.

Define the set of functions $\mathcal{F} = \Big\{f(z_i, A) \,:\, A \in \mathcal{A} \text{ for } i = 1, ..., n\Big\}$ for $f : \mathcal{Z} \times \Theta \times \mathcal{A} \to \mathbb{R}$ such that $f(z_i, A) = g^{(k)}(z_i, \theta)\mathbb{I}\{z_i \in A\}$. Therefore $\mathcal{F}$ is a class of real-valued functions from

$\mathcal{Z} \to \big[-C_g, C_g\big]$ from Assumption 2.1 (ii). If $\mathcal{A}$ is a class of sets with VC dimension v, then by Lemma C.1 with $h(z_i) = g^{(k)}(z_i, \theta_0)$, $\mathcal{F}$ is a VC-subgraph with VC-dimension $2v+1$. The problem of picking the best set $A \in \mathcal{A}$ is then equivalent to picking the best function f in the class $\mathcal{F}$. I.e. for $f_n(z) = n^{-1}\sum_{i=1}^n f(z_i)$ and $f_n(z') = n^{-1}\sum_{i=1}^n f(z_i')$,

$$\mathbb{P}_{2n}\Big(\sup_{A\in\mathcal{A}}\Big|f_n(A) - f_n'(A)\Big| > \frac{\delta}{2}\Big) = \mathbb{P}_{2n}\Big(\sup_{f\in\mathcal{F}}\Big|f_n(z) - f_n'(z)\Big| > \frac{\delta}{2}\Big).$$

Then by Lemma C.4,

$$\mathbb{P}_{2n}\Big(\sup_{f\in\mathcal{F}}\Big|f_n(z) - f_n'(z)\Big| > \frac{\delta}{2}\Big) \leq 2 \times \mathbb{P}_n\Big(\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^n \sigma_i f(z_i)\Big| > \frac{\delta}{4}\Big).$$

Now fix and condition on $z = (z_1, ..., z_n)$,

$$\mathbb{P}_n\Big(\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^n \sigma_i f(z_i)\Big| > \frac{\delta}{4}\Big) = \mathbb{E}_z\Big[\mathbb{P}_n\Big(\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^n \sigma_i f(z_i)\Big| > \frac{\delta}{4}\Big|z_1, ..., z_n\Big)\Big].$$

Let $\mathcal{F}_{\frac{\delta}{8}}$ be a $\frac{\delta}{8}$-cover of $\mathcal{F}$ with respect to the metric d_1 ($d_1(f,q) = \frac{1}{n}\sum_{i=1}^n |f(z_i) - q(z_i)|$), defined on the points $z_1, ..., z_n$ of minimal cardinality $\mathcal{N}(\frac{\delta}{8}, \mathcal{F}_{|z}, d_1)$. Then for any $f \in \mathcal{F}$, there exists $q \in \mathcal{F}_{\frac{\delta}{8}}$ such that $\frac{1}{n}\sum_{i=1}^n |f(z_i) - q(z_i)| < \frac{\delta}{8}$.

For $\Big|\frac{1}{n}\sum_{i=1}^n \sigma_i f(z_i)\Big| > \frac{\delta}{4}$, note that $\frac{1}{n}\sum_{i=1}^n \sigma_i f(z_i) > \frac{\delta}{4}$ or $\frac{1}{n}\sum_{i=1}^n \sigma_i f(z_i) < -\frac{\delta}{4}$. By definition of the cover,

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \sigma_i f(z_i) &< \frac{1}{n}\sum_{i=1}^n \sigma_i q(z_i) + \frac{\delta}{8} \\
\frac{1}{n}\sum_{i=1}^n \sigma_i f(z_i) &> \frac{1}{n}\sum_{i=1}^n \sigma_i q(z_i) - \frac{\delta}{8}.
\end{aligned}$$

Therefore,

$$\mathbb{P}_n\Big(\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^n \sigma_i f(z_i)\Big| > \frac{\delta}{4}\Big|z_1, ..., z_n\Big) \leq \mathbb{P}_n\Big(\max_{q\in\mathcal{F}_{\frac{\delta}{8}}}\Big|\frac{1}{n}\sum_{i=1}^n \sigma_i q(z_i)\Big| > \frac{\delta}{8}\Big|z_1, ..., z_n\Big).$$

By the union bound,

$$\begin{aligned}\mathbb{P}_n\Big(\max_{q\in\mathcal{F}_{\frac{\delta}{8}}}\Big|\frac{1}{n}\sum_{i=1}^n\sigma_i q(z_i)\Big|>\frac{\delta}{8}\Big|z_1,...,z_n\Big) &\leq \mathbb{P}_n\Big(\bigcup_{q\in\mathcal{F}_{\frac{\delta}{8}}}\Big\{\Big|\frac{1}{n}\sum_{i=1}^n\sigma_i q(z_i)\Big|>\frac{\delta}{8}\Big\}\Big|z_1,...,z_n\Big)\\ &\leq \sum_{q\in\mathcal{F}_{\frac{\delta}{8}}}\mathbb{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^n\sigma_i q(z_i)\Big|>\frac{\delta}{8}\Big|z_1,...,z_n\Big)\\ &\leq \big|\mathcal{F}_{\frac{\delta}{8}}\big|\times\mathbb{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^n\sigma_i q(z_i)\Big|>\frac{\delta}{8}\Big|z_1,...,z_n\Big).\end{aligned}$$

By Lemma B.3,

$$\mathbb{P}_n\Big(\Big|\sum_{i=1}^n\sigma_i q(z_i)\Big|>\frac{n\delta}{8}\Big|z_1,...,z_n\Big)\leq 2\exp\Big(-\frac{\delta^2 n}{128C_g^2}\Big).$$

Also, $\mathbb{E}_z(\big|\mathcal{F}_{\frac{\delta}{8}}\big|)=\mathbb{E}_z(\mathcal{N}(\frac{\delta}{8},\mathcal{F}_{|z},d_1))\leq\mathcal{N}_1(\frac{\delta}{8},\mathcal{F},n)\leq\mathcal{N}_\infty(\frac{\delta}{8},\mathcal{F},n)$. Then by Lemmata C.1 and C.2, for $n>2v+1$,

$$\mathcal{N}_\infty(\frac{\delta}{8},\mathcal{F},n)\leq\Big(\frac{16enC_g}{(2v+1)\delta}\Big)^{2v+1}.$$

Overall,

$$\mathbb{P}_n\Big(\sup_{A\in\mathcal{A}}\big|f_n(A)-\mathbb{E}f(z,A)\big|>\delta\Big)\leq 8\Big(\frac{16enC_g}{(2v+1)\delta}\Big)^{2v+1}\exp\Big(-\frac{\delta^2 n}{128C_g^2}\Big)$$

for $n>\max\{2v+1,\,8C_g^2/\delta^2\}$. □

D Bounds of GEL Quantities

The following bounds for various GEL quantities are useful. Many of the proofs make use of the property that the Euclidean norm is bounded above by the ℓ_1 norm.

Lemma D.1. *Under Assumption 2.1,* **(i)** $\sup_{\theta\in\Theta}||g_n(\theta)||\leq d_g C_g$; **(ii)** $\sup_{\theta\in\Theta}||\mathbb{E}[g(z,\theta)]||\leq d_g C_g$;
(iii) $\sup_{\theta\in\Theta}||\frac{1}{n}\sum_{i=1}^n g_i(\theta)\mathbb{I}\{z_i\in A\}||\leq d_g C_g$; **(iv)** $\sup_{A\in\mathcal{A}}||\mathbb{E}[g(z,\theta_0)\mathbb{I}\{z\in A\}]||\leq d_g C_{B,\mathcal{A}}$; **(v)** $\sup_{\theta\in\Theta}\big|\big|\frac{1}{n}\sum_{i=1}^n G_i(\theta)\big|\big|\leq d_g C_G$.

Proof. Write

$$\begin{aligned}\sup_{\theta\in\Theta}||g_n(\theta)|| &\leq \sup_{\theta\in\Theta}\sum_{k=1}^{d_g}\Big|\frac{1}{n}\sum_{i=1}^{n}g^{(k)}(z_i,\theta)\Big| \\ &\leq \sum_{k=1}^{d_g}\sup_{z\in\mathcal{Z}}\sup_{\theta\in\Theta}|g^{(k)}(z,\theta)| \\ &\leq d_g C_g\end{aligned}$$

by Assumption 2.1(ii).

Similarly for (ii),

$$\begin{aligned}\sup_{\theta\in\Theta}||\mathbb{E}[g(z,\theta_0)]|| &\leq \sup_{\theta\in\Theta}\sum_{k=1}^{d_g}|\mathbb{E}[g^{(k)}(z,\theta_0)]| \\ &\leq \sum_{k=1}^{d_g}\mathbb{E}\Big[\sup_{z\in\mathcal{Z}}\sup_{\theta\in\Theta}|g^{(k)}(z,\theta_0)|\Big] \\ &\leq d_g C_g\end{aligned}$$

by Assumption 2.1(ii).

For (iii),

$$\begin{aligned}\sup_{\theta\in\Theta}||\frac{1}{n}\sum_{i=1}^{n}g_i(\theta)\mathbb{I}\{z_i\in A\}|| &\leq \sup_{\theta\in\Theta}\sum_{k=1}^{d_g}|\frac{1}{n}\sum_{i=1}^{n}g_i^{(k)}(\theta)\mathbb{I}\{z_i\in A\}| \\ &\leq \sum_{k=1}^{d_g}\sup_{z\in\mathcal{Z}}\sup_{\theta\in\Theta}|g^{(k)}(z,\theta)| \\ &\leq d_g C_g\end{aligned}$$

by Assumption 2.1(ii).

For (iv),

$$\begin{aligned}\sup_{A\in\mathcal{A}}||\mathbb{E}[g(z,\theta_0)\mathbb{I}\{z\in A\}|| &\leq \sum_{k=1}^{d_g}\sup_{A\in\mathcal{A}}|\mathbb{E}[g^{(k)}(z,\theta_0)\mathbb{I}\{z\in A\}]| \\ &\leq d_g C_{B,\mathcal{A}}\end{aligned}$$

by Assumption 2.2(viii).

Finally, for (v),

$$\begin{aligned}\sup_{\theta\in\Theta}\Big|\Big|\frac{1}{n}\sum_{i=1}^{n}G_i(\theta)\Big|\Big|_1 &\leq \max_{1\leq j\leq d_\theta}\sum_{k=1}^{d_g}\frac{1}{n}\sum_{i=1}^{n}\sup_{\theta\in\Theta}\Big|\frac{g^{(k)}(z_i,\theta)}{\partial\theta_j}\Big| \\ &\leq \sum_{k=1}^{d_g}\sup_{z\in\mathcal{Z}}\sup_{\theta\in\Theta}\Big|\frac{g^{(k)}(z,\theta)}{\partial\theta_j}\Big| \\ &\leq d_g C_G\end{aligned}$$

by Assumption 2.1(iii). □

Lemma D.2. *Under Assumption 2.1(iv),* **(i)** $\sup_{\theta\in\Theta}||\Omega^{-1}(\theta)|| \leq 1/\delta_{\Omega,\min}$*;* **(ii)** $\sup_{\theta\in\Theta}||\Omega_n^{-1}(\theta)|| \leq \frac{1}{\delta_{\Omega,\min}} + \sup_{\theta\in\Theta}||\Omega_n^{-1}(\theta)-\Omega^{-1}(\theta)||_1$.

Proof. By Assumption 2.1(iv), $\Omega(\theta)$ has minimum eigenvalue $\delta_{\Omega,\min}$. (i) immediately follows. For (ii),

$$\begin{aligned}\sup_{\theta\in\Theta}||\Omega_n^{-1}(\theta)|| &\leq \sup_{\theta\in\Theta}||\Omega^{-1}(\theta)|| + \sup_{\theta\in\Theta}||\Omega_n^{-1}(\theta)-\Omega^{-1}(\theta)|| \\ &\leq \frac{1}{\delta_{\Omega,\min}} + \sup_{\theta\in\Theta}||\Omega_n^{-1}(\theta)-\Omega^{-1}(\theta)|| \\ &\leq \frac{1}{\delta_{\Omega,\min}} + \sup_{\theta\in\Theta}||\Omega_n^{-1}(\theta)-\Omega^{-1}(\theta)||_1\end{aligned}$$

where the second inequality follows from Part (i). □